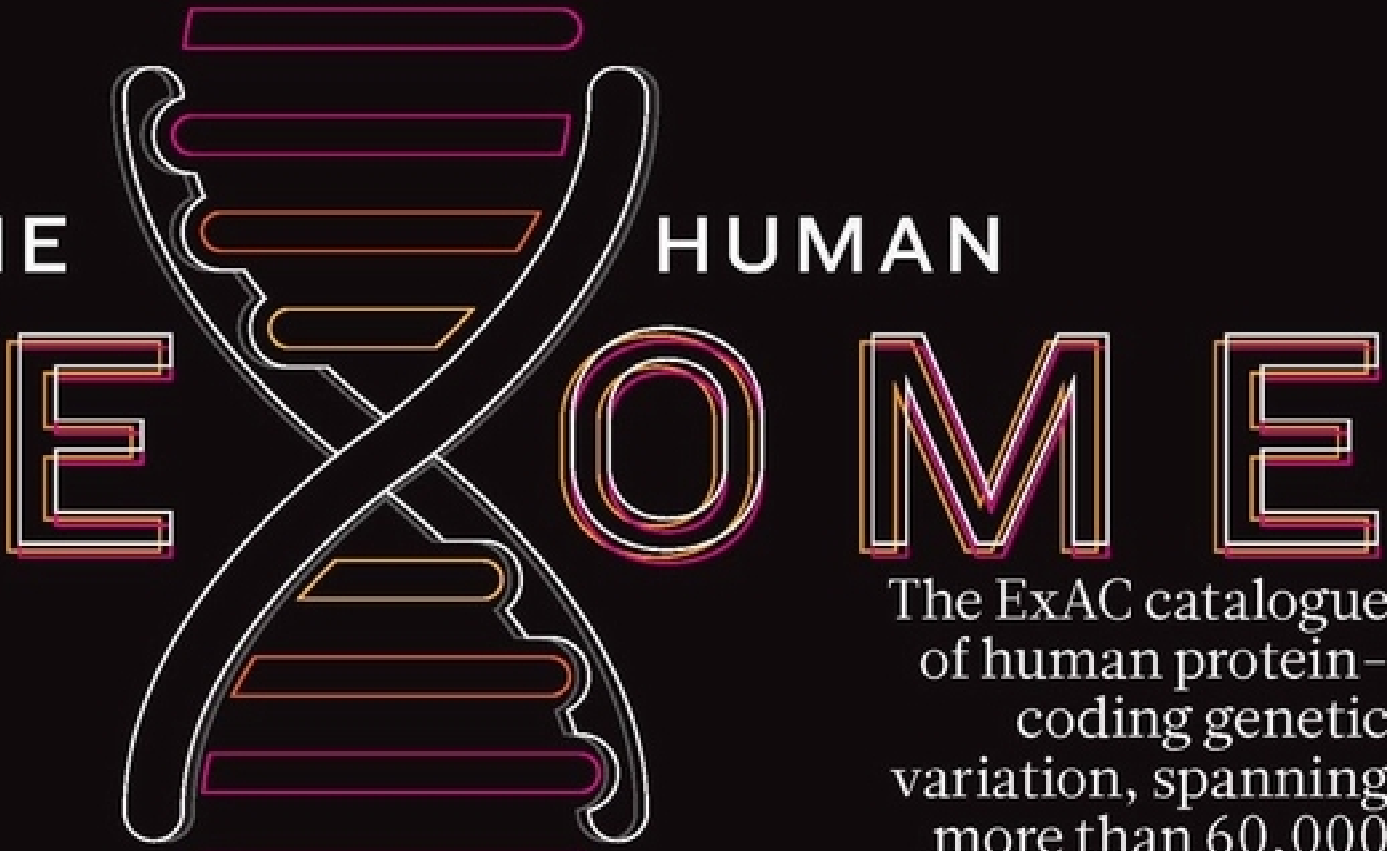
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

SPECIAL ISSUE
Polymers and plastics



THE HUMAN EXOME

The ExAC catalogue of human protein-coding genetic variation, spanning more than 60,000 samples **PAGES 277 & 285**

TOXICOLOGY

GIVE CHEMISTS A CHANCE

New testing protocols would stimulate invention

PAGE 269

NEUROSCIENCE

THROW THE SLEEP SWITCH

Dopamine signals at the waking/sleeping transition

PAGES 278 & 333

STRUCTURAL BIOLOGY

SEEING IS BELIEVING

Coping with a smorgasbord of imaging technologies

PAGE 361

NATURE.COM/NATURE

18 August 2016 £10

Vol. 536, No. 7616

THIS WEEK

EDITORIALS

PLANETARY SCIENCES Don't misdefine — or underplay — the Anthropocene **p.251**

EARTH SCIENCES Ancient chunk of crust lurks in Mediterranean **p.252**

ANIMAL WELFARE NIH chimps have date set for retirement home **p.254**

Rare rewards

A catalogue of genetic information from some 60,000 people reveals unexpected surprises — and highlights the need to make genomic data publicly accessible to aid studies of rare diseases.

More than one million people have now had their genome sequenced, or its protein-coding regions (the exome). The hope is that this information can be shared and linked to phenotype — specifically, disease — and improve medical care. An obstacle is that only a small fraction of these data are publicly available.

In an important step, we report this week the first publication from the Exome Aggregation Consortium (ExAC), which has generated the largest catalogue so far of variation in human protein-coding regions (see page 285). It aggregates sequence data from some 60,000 people. Most importantly, it puts the information in a publicly accessible database that is already a crucial resource (http://exac.broadinstitute.org).

There are challenges in sharing such data sets — the project scientists deserve credit for making this one open access. Its scale offers insight into rare genetic variation across populations. It identifies more than 7.4 million (mostly new) variants at high confidence, and documents rare mutations that independently emerged, providing the first estimate of the frequency of their recurrence. And it finds 3,230 genes that show nearly no cases of loss of function. More than two-thirds have not been linked to disease, which points to how much we have yet to understand.

The study also raises concern about how genetic variants have been linked to rare disease. The average ExAC participant has some 54 variants previously classified as causal for a rare disorder; many show up at an implausibly high frequency, suggesting that they were incorrectly classified. The authors review evidence for 192 variants reported earlier to cause rare Mendelian disorders and found at a high frequency by ExAC, and uncover support for pathogenicity for only 9. The implications are broad: these variant data already guide diagnoses and treatment (see, for example, E. V. Minikel *et al. Sci. Transl. Med.* **8,** 322ra9; 2016 and R. Walsh *et al. Genet. Med.* http://dx.doi.org/10.1038/gim.2016.90; 2016).

These findings show that researchers and clinicians must carefully evaluate published results on rare genetic disorders. And it demonstrates the need to filter variants seen in sequence data, using the ExAC data set and other reference tools — a practice widely adopted in genomics.

The ExAC project plans to grow over the next year to include 120,000 exome and 20,000 whole-genome sequences. It relies on the willingness of large research consortia to cooperate, and highlights the huge value of sharing, aggregation and harmonization of genomic data. This is also true for patient variants — there is a need for databases that provide greater confidence in variant interpretation, such as the US National Center for Biotechnology Information's ClinVar database.

Improving clinical genetics will need continued investment in such databases, more contributions from clinical labs, researchers and clinicians, expanding human genetic-reference panels and work to link these to phenotype data. This often involves re-contacting volunteers and donors; it will be trialled with an ExAC data subset where consents allow.

More broadly, enabling the sharing of linked genetic and clinical data in ways that do not violate privacy requires fresh thinking in regulation and ethics. The US National Institutes of Health and the Global Alliance for Genomics and Health have begun to tackle this; others should follow. The ExAC study highlights the potential rewards. ■

Evo-devo CRISPR

Modern gene-editing tools are helping to unpick the origins of evolutionary adaptations.

Some call it the chickenosaurus, others the dino-chicken. Whatever you term the proposal to transform a chicken into a creature more like its dinosaur ancestor, it is having its scientific moment.

Researchers have succeeded in making the feet, limbs and face of unhatched chicks a bit more like those of the creatures' 150-million-year-old ancestors by tinkering with the molecular pathways that forge these structures. The goal is to comprehend the molecular events responsible for one of the most awe-inspiring transitions in the fossil record.

The field of evolutionary developmental biology — evo-devo — is full of such creations: from mice with longer, bat-like limbs to fruit flies with torsos segmented like beetles'. But until now, the brute tools used to create these creatures have been imperfect.

This is about to change. In a paper published online on 17 August, a team used CRISPR–Cas9 to inactivate the genes involved in zebrafish development, resulting in fin tips more like the feet and digits of land vertebrates (T. Nakamura *et al. Nature* http://dx.doi.org/10.1038/nature19322; 2016). Other recent CRISPR experiments have tinkered with butterflies to learn how they see more colours than flies do, and done away with crustaceans' claws to understand the origin of these specialized appendages.

So far, the edits have tended to simply inactivate genes. But evo-devo scientists will soon start swapping genes between distantly related beasts to learn the origins of adaptations such as multicellularity and the anus, to name but two problems troubling the field. Our ability to access and analyse ancient DNA means that we can now insert genes from extinct animals into the genomes of their living relatives.

These sorts of experiments could draw evo-devo fancifully close to de-extinction, the quest to resurrect woolly mammoths and other long-dead animals. But every upturned urinal is not a Dadaist masterwork, and the idea behind the experiments is what matters. These 'hopeful CRISPR monsters' could confirm or reject decades-old theories about key events in evolution, and help us to come up with new ones.

Just think of what we could learn from a bona fide dino-chicken. ■

WORLD VIEW *A personal take on events*

Define the Anthropocene in terms of the whole Earth

Researchers must consider human impacts on entire Earth systems and not get trapped in discipline-specific definitions, says **Clive Hamilton***.*

Do we live in the Anthropocene? Officially, not yet — although the debate about whether to declare a new geological epoch will resurface later this month at the International Geological Congress in Cape Town, South Africa. The concept of the Anthropocene has become well known and is much discussed, but often in a way that undermines the seriousness of the issue.

The Anthropocene was conceived by Earth-system scientists to capture the very recent rupture in Earth's history arising from the impact of human activity on the Earth system as a whole. Read that again. Take special note of the phrases 'very recent rupture' and 'the Earth system as a whole'. Understanding the Anthropocene, and what humanity now confronts, depends on a firm grasp of these concepts, and that they arise from the new discipline of Earth-system science. Earth-system science takes an integrated approach, so that climate change affects the functioning of not just the atmosphere, but also the hydrosphere, the cryosphere, the biosphere and even the lithosphere. (Arguably, anthropogenic climate change is more an oceanic than an atmospheric phenomenon.)

In the canonical statement of the Anthropocene, the proposed new division in the geological timescale is defined by the observation that the "human imprint on the global environment has now become so large and active that it rivals some of the great forces of Nature in its impact on the functioning of the Earth system" (W. Steffen *et al. Phil. Trans. R. Soc. A* **369,** 842–867; 2011). As such, the Anthropocene cannot be defined merely by the broadening impact of people on the environment and natural world, which just extends what we have done for centuries or millennia.

Yet this is how many scientists are trying to define it. And this is because much discussion of the Anthropocene — its essential idea, its causes, its timing — is bedevilled by readings through old disciplinary lenses, which don't account for the true implications of humankind taking the planet into a new epoch.

THE ANTHROPOCENE **CANNOT** BE DEFINED MERELY BY THE **BROADENING** IMPACT OF PEOPLE ON THE ENVIRONMENT AND NATURAL **WORLD.**

Probably the most obvious example of scientific misinterpretation of the Anthropocene is the debate about its starting date. Discussions on rival starting dates may seem to have scientific merit, but they distort and dilute the message and the implications of the Anthropocene.

The original suggested onset was the end of the eighteenth century, when the European industrial revolution's large-scale coal-burning triggered rising concentrations of carbon dioxide in the atmosphere. More recently, members of the Anthropocene Working Group have proposed — I think correctly — 1945 as an unambiguous beginning for people causing a shift in the functioning of the Earth system.

But peering through the narrow lens of landscape ecology, others have interpreted the new geological epoch as another name for the continued impact of people on the terrestrial biosphere. Changing vegetation and landscapes may bear the hallmarks of human behaviour, but these cannot have sufficient impact on the Earth system to bring about a new geological epoch.

Others misconstrue the question from the outset and argue that the Anthropocene's starting date depends on when human societies first began to play a significant part in shaping Earth's ecosystems. The very last letter, the 's' in ecosystems, gives it away. The Anthropocene began not when humans first played a significant part in those, but when they first changed the functioning of the Earth system. With a similar sleight of hand, others insert archaeology into the debate, so that the Anthropocene can be traced to the first domestication of plants and animals some 10,000 years ago. And some go further still and insist the Anthropocene is the most recent phase of a process that started 50,000 years ago with human geographic expansion.

Geographers and soil scientists have also claimed the Anthropocene for themselves. The start of the new epoch is 1610, the geographers say, based on a complex narrative covering the colonization of South America, introduced diseases, depopulation, forest regrowth, transcontinental trade, species exchange and pollen counts. Soil scientists put the date more than 1,000 years earlier, with evidence for anthropogenic modification of soils.

One thing all these misreadings of the Anthropocene have in common is that they divorce it from modern industrialization and the burning of fossil fuels. In this way, the Anthropocene no longer represents a rupture in Earth history but is a continuation of the kind of impact people have always had. This thereby renders it benign, and the serious and distinct threat of climate change becomes just another human influence.

That so many scientists, often publishing in prestigious journals, can misconstrue the definition of the Anthropocene as nothing more than a measure of the human footprint on the landscape is a sign of how far Earth-system science has to go to change the way many people think about the planet. The new geological epoch does not concern soils, the landscape or the environment, except inasmuch as they are changed as part of a massive shock to the functioning of Earth as a whole.

Some scientists even write: "Welcome to the Anthropocene." At first I thought they were being ironic, but now I see they are not. And that's scary. The idea of the Anthropocene is not welcoming. It should frighten us. And scientists should present it as such. ■

Clive Hamilton *is professor of public ethics at Charles Sturt University in Canberra, Australia, and author of* Defiant Earth: The Fate of Humans in the Anthropocene, *to be published next year.*
e-mail: mail@clivehamilton.com

RESEARCH HIGHLIGHTS

Selections from the scientific literature

AGRICULTURAL ECOLOGY

Pesticide link to wild-bee declines

A class of pesticide called neonicotinoids has been associated with the decline of wild-bee species across the United Kingdom.

Small and short-term studies have shown that the chemicals — which were first used widely in the country in 2002, before being placed under a 2-year moratorium by the European Union in 2013 — can harm bee reproduction. To look for long-term effects at the population level, Ben Woodcock at the NERC Centre for Ecology and Hydrology in Wallingford, UK, and his colleagues compared maps of pesticide use on oilseed rape (canola) crops with surveys of 62 wild-bee species across the United Kingdom from 1994 to 2011. They found correlations between neonicotinoid exposure and population declines in bees that forage on the crops, and even in some that don't.

The team estimates that the chemicals are linked to population losses of more than 10% for 24 bee species.

Nature Commun. **7, 12459 (2016)**

SUE CUNNINGHAM PHOTOGRAPHIC/ALAMY

CLIMATE CHANGE

Warming drives down lake life

Rising temperatures have lowered fish numbers in one of Africa's great lakes, threatening food sources vital to local people.

Andrew Cohen at the University of Arizona in Tucson and his colleagues analysed sediments and fossils from Lake Tanganyika (pictured) to infer water temperatures and estimate species abundance going back over 1,500 years. They found population declines in fishes, molluscs and plankton that pre-dated commercial fishing, but correlated with sustained warming and falling algal production during the past 150 years.

Warming reduces the mixing of nutrient-rich deeper waters with oxygen-rich shallow waters, limiting the growth of plankton — an important food source for many fishes. Reduced mixing also lowers the area of oxygenated water at the bottom of the lake, threatening numerous fish and invertebrate species.

Proc. Natl Acad. Sci. USA **http://doi.org/bnqk (2016)**

BIODIVERSITY

New lizards under threat

Recently discovered lizard species tend to be smaller, are more often nocturnal and are at greater risk of extinction than those described previously.

Scientists have been identifying new lizard species at an astonishing rate — with a more than 30% increase in species number recorded since 2000. To find out what these animals have in common, Shai Meiri at Tel Aviv University in Israel studied data on the biology and geography of all 6,321 lizard species known in mid-2015. He found that species described this century tended to be small and to have limited geographical ranges — explaining why they remained undiscovered for so long. Nearly 40% of these lizards were geckos, and 37% were nocturnal.

New species were more likely to have declining populations and face extinction, meaning that many species may be lost soon after — or perhaps even before — being described, Meiri warns.

J. Zool. **299, 251–261 (2016)**

CELL BIOLOGY

CRISPR switches cell types

By activating a suite of genes using the gene-targeting tool CRISPR–Cas9, researchers have turned connective-tissue cells called fibroblasts directly into neurons.

Directly reprogramming cells from one identity to another could one day provide abundant material for disease research or therapies. But scientists face a technical challenge — keeping genes required for the new identity switched on for a lengthy period of time. To resolve this, Charles Gersbach at Duke University in Durham, North Carolina, and his colleagues used a CRISPR–Cas9-based system to activate three genes, converting mouse embryonic fibroblasts into neuronal cells and sustaining gene activation throughout the process.

The technique could provide a way to reprogram cells without having to insert genes into the genome.

Cell Stem Cell **http://doi.org/bn22 (2016)**

JAMES ST CLAIR

PLANETARY SCIENCE

Methane-filled canyons on Titan

The surface of Saturn's largest moon is etched with canyons that are flooded with liquid hydrocarbons, according to data from NASA's Cassini spacecraft.

Valerio Poggiali of the Sapienza University of Rome and his team used radar aboard Cassini to measure elevations on Titan and map out a network of steep-sided, narrow channels called Vid Flumina. Some of the canyons are up to 570 metres deep. Titan has low average temperatures of −179 °C, so it previously wasn't clear whether the dark material in these canyons was ice. However, the scientists found that liquid methane flows through the channels and into the northern sea, Ligeia Mare.

Other than Earth, Titan is the only planetary body in the Solar System that has active erosion caused by liquid on its surface.

Geophys. Res. Lett. **http://doi.org/bn2p (2016)**

NANOMATERIALS

Sunlight helps to purify water

Nanometre-thin films can harvest natural light and use it to rapidly disinfect water.

Sunlight offers a useful means of purifying water, particularly in countries that lack reliable energy sources. Ultraviolet light is widely used to kill microbes, but accounts for only 4% of the solar spectrum. Yi Cui and his colleagues at Stanford University in California have developed a film — comprising vertically aligned layers of molybdenum disulfide — that captures visible light, taking advantage of about 50% of the total solar energy. Light causes the films to generate reactive oxygen molecules, which kill water-borne pathogens.

Placing the film in water containing *Escherichia coli* and exposing it to light led to near-total disinfection in 20 minutes. Previous systems needed 30 to 60 minutes.

Nature Nanotechnol. **http://dx.doi.org/10.1038/nnano.2016.138 (2016)**

GEOPHYSICS

Ancient sea floor preserved

The eastern Mediterranean Sea contains a surprisingly ancient chunk of oceanic crust, which is probably helping to shape the region's geology today.

The shifting of Earth's crustal plates has destroyed most oceanic rock older than about 200 million years. Roi Granot at Ben-Gurion University of the Negev in Beer-Sheva, Israel, investigated hints that the Herodotus Basin in the eastern Mediterranean might be older than that. Data from ship-towed instruments revealed long stripes of alternating magnetism on the Herodotus sea floor — a characteristic suggesting that it is oceanic, rather than continental, crust. The geometry of the stripes indicates that the crust dates back some 340 million years.

Earthquakes are a frequent occurrence on the sea floor where this relatively strong ancient crust meets weaker continental crust to the east.

Nature Geosci. **http://dx.doi.org/10.1038/ngeo2784 (2016)**

ANIMAL COGNITION

Crafty crows bend their tools

Creating bent tools to fish for food in holes and crevices seems to come naturally to a species of crow.

In 2002, a captive New Caledonian crow (*Corvus moneduloides*) called Betty astonished scientists by bending straight pieces of wire into hooked tools to access out-of-reach food. But recent field experiments by Christian Rutz and his colleagues at the University of St Andrews, UK, show that bending is used by wild New Caledonian crows, too (**pictured**). More than half of the 18 wild-caught crows in the study bent sticks during routine tool manufacture, using methods similar to those used by Betty to handle wire. Most birds stood on the sticks and pulled the tip up.

This discovery suggests that bending may have been part of Betty's natural tool-crafting repertoire, rather than a smart invention.

R. Soc. Open Sci. 3, **160439 (2016)**

ZOOLOGY

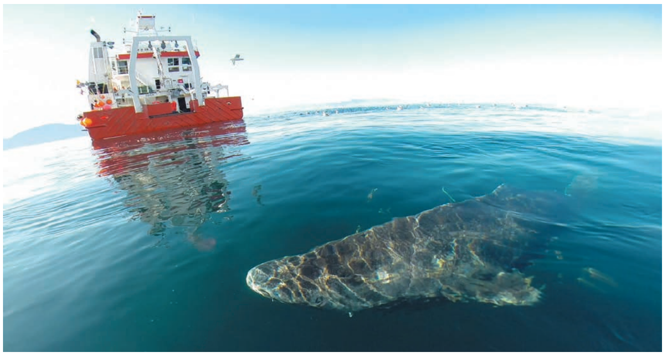
Sharks live for centuries

A shark species found in Arctic seas may live for up to 400 years, making it the longest-lived vertebrate known.

Julius Nielsen at the University of Copenhagen and his colleagues estimated the ages of 28 female Greenland sharks (*Somniosus microcephalus*; **pictured**), by radiocarbon dating the nuclei in the animals' eye lens. They concluded that the animals have a lifespan of at least 272 years, and that females don't reach sexual maturity until they are more than 100 years old.

The findings raise concerns about Greenland shark conservation, because a species that takes so long to begin reproducing could be at risk of being over-exploited by fisheries. The animal is also often inadvertently captured in nets cast for other species.

Science 353, **702–704 (2016)**



JULIUS NIELSEN

ATOMIC PHYSICS

Proton-size puzzle deepens

Atomic measurements add weight to recent work suggesting that the proton is significantly smaller than previously thought.

In 2010, researchers studied muonic hydrogen (in which the electron is replaced with a muon, a bigger particle that is also negatively charged), which allowed them to measure the nuclear radius much more accurately than is possible with ordinary hydrogen. The proton size they found was smaller than expected from previous measurements. To confirm the results, the same team, led by Randolf Pohl of the Max-Planck Institute for Quantum Optics in Garching, Germany, studied the nucleus of muonic deuterium, which contains a proton and a neutron. They calculated that the proton radius is about 5% smaller than previously measured — a similar result to that of 2010.

Several laboratories are redoing the measurements in ordinary hydrogen atoms to try to resolve the contradiction.

Science 353, **669–673 (2016)**

➲ NATURE.COM

For the latest research published by *Nature* visit:

www.nature.com/latestresearch

SEVEN DAYS *The news in brief*

SUNDAY ALAMBA/AP

Polio returns to Nigeria after two years

Nigeria has recorded its first cases of wild poliovirus in more than two years — a setback to the global campaign to eradicate the disease. The country will now start emergency vaccination campaigns to hold back the virus's spread. Nigeria's government found two children in the northeastern Borno state who had been paralysed by polio in July, the World Health Organization announced on 11 August. The country had been on the brink of wiping out polio; its last recorded case had been in July 2014. See go.nature.com/2bccf92 for more.

FUNDING

Grants guaranteed

The UK government has announced that it will guarantee existing research grants from the European Union's Horizon 2020 funding programme after the country leaves the EU. UK researchers receive billions of euros from the €75-billion (US$84-billion) programme. June's referendum vote to leave the EU left British scientists worried that funding for projects spanning several years could be taken away. But on 13 August, the government said that it would cover any shortfalls in these grants, provided that an organization has bid for them before the United Kingdom leaves the EU (a date that has not yet been fixed). See go.nature.com/2aoa7gi for more.

US finds Zika funds

The US government will put US$81 million towards the study of potential vaccines against the Zika virus, the US Department of Health and Human Services announced on 11 August. The money, which is to be reallocated from other projects, will be divided between the National Institutes of Health (NIH), which will receive $34 million, and the Biomedical Advanced Research and Development Authority, which will receive $47 million. Even so, the NIH estimates that it will need another $196 million for Zika vaccine research in the 2017 fiscal year.

RESEARCH

Solar-storm war risk

A study has revealed for the first time how a solar storm in May 1967 nearly caused the US military to launch planes for war — until Air Force researchers realized that ballistic early-warning radars were being jammed by energetic solar particles, and not by the Soviet Union (D. J. Knipp *et al. Space Weather* http://doi.org/bn5x; 2016). Since then, the Department of Defense has continued to invest heavily in space-weather forecasting. A team led by Delores Knipp, a space physicist at the University of Colorado Boulder, disclosed the story at a meeting in Boulder on 10 August.

EVENTS

Quantum satellite

China launched the world's first quantum satellite on 16 August. The Quantum Experiments at Space Scale (QUESS) mission, which lifted off from the Jiuquan Satellite Launch Center in northern China, successfully entered orbit at an altitude of 500 kilometres. During its two-year mission, QUESS will test the limits of the quantum phenomenon known as entanglement by observing whether entangled photons remain linked across 1,200 kilometres, eight times the distance so far achieved in free space. It will also test ways to 'teleport' information between the satellite and Earth using entangled photons.

ChemRxiv coming

The American Chemical Society announced on 10 August that it wants to establish a preprint site for chemistry called ChemRxiv. The site would allow chemists to share early results and data online before publication. The repository would follow the physics preprint server arXiv and the bioRxiv for biologists. Chemists have historically been reluctant to publicly share manuscripts before peer review. One reason is that some major journals in the discipline discourage posting work online before submission. See go.nature.com/2bn7clg for more.

Corruption case

India's Central Bureau of Investigation has charged the former chair of the Indian Space Research Organisation (ISRO), G. Madhavan Nair, and seven leading officials in a corruption case. The group includes former employees of the ISRO and its commercial arm Antrix, and officials of the satellite company Devas Multimedia in Bangalore. They have been accused of cheating, corruption and conspiracy to make financial gains by abusing their positions in a deal between Antrix and Devas. In 2011, Antrix scrapped a 2005 contract with Devas, citing security concerns amid allegations of corruption. Nair, who was head of Antrix as well as the ISRO, signed off the deal. Devas took the case to the courts, and two international arbitration proceedings have ruled in the company's favour.

POLICY

Chimp retirement

The US National Institutes of Health has released a long-awaited plan to move its research chimpanzees into permanent retirement. Under the schedule, announced by the agency on 11 August, all 360 of the agency's chimpanzees will be moved to the federally funded Chimp Haven sanctuary (**pictured**) in Louisiana by 2026. The chimps, which are living in research centres in Texas and New Mexico, will be moved in small groups to keep families and social circles together. Chimp Haven currently has space for only about 75 more animals, and is constructing space for an extra 100; deaths in the ageing population are likely to make further room.

MELANIE STETSON FREEMAN/CHRISTIAN SCIENCE MONITOR/GETTY

More marijuana

The United States is making it easier to access marijuana for research purposes. Scientists have long been able to obtain the drug from only one source — the University of Mississippi in Oxford. But in an unexpected move, the US Drug Enforcement Administration announced on 11 August that it will allow any institution to apply for permission to grow marijuana for research. The agency says that the change is motivated by a high demand from scientists and a desire to encourage research on the drug. Research on marijuana typically focuses on cannabinoids, compounds found in the drug that may alleviate chronic pain and mitigate the effects of neurological disorders.

PEOPLE

Harvard chief

Leading stem-cell scientist George Daley was announced as the new dean of Harvard Medical School in Boston, Massachusetts, on 9 August. Daley is a long-time faculty member at the school and its affiliated hospitals. He is a former president of the International Society for Stem Cell Research, which has established guidelines for the use of embryonic stem cells, and most recently, for the gene editing of human germline cells.

FACILITIES

US space projects

A report on the United States' biggest astronomy and astrophysics programmes has affirmed the country's support for a European Space Agency-led gravitational-wave observatory. Physicists announced the first detection of gravitational waves in February; since then, the proposed space-based observatory has garnered renewed US interest. The report, released on 15 August, is a 'midterm' assessment of decadal funding priorities laid out in 2010 for agencies including NASA and the US National Science Foundation. The Wide-Field Infrared Survey Telescope, which was widely favoured in 2010, is still slated for launch in the mid-2020s.

Chinese sea lab

China is to establish a South China Sea marine-research laboratory on Hainan island. The State Key Laboratory of Marine Resources Utilization in the South China Sea, which is due to open in November with an initial term of five years, will survey mineral and marine resources in the region. The announcement came shortly after a decision by the Permanent Court of Arbitration in The Hague, the Netherlands, that censured China for transgressing international law in its attempt to claim nearly all of the South China Sea. China rejected the ruling. According to sources in state press, the laboratory will help China to "safeguard our nation's rights".

COMING UP

21–26 AUGUST
Scientists gather in Melbourne, Australia, for the International Congress of Immunology.
ici2016.org

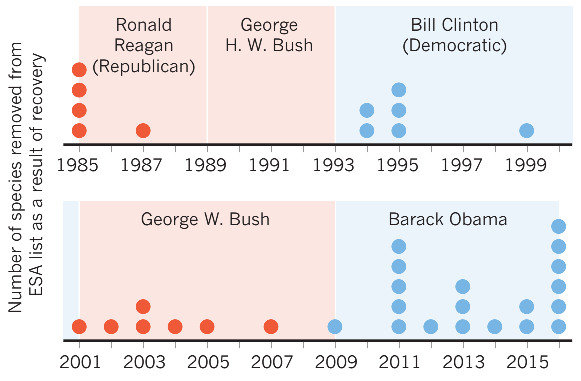
24 AUGUST
The US Animal Welfare Act, the first federal law to regulate the use of animals in research, turns 50.

TREND WATCH

More species protected by the US Endangered Species Act have recovered during President Barack Obama's administration than under all other presidents combined, the US Department of the Interior announced on 11 August. Under Obama, 19 species have recovered and been delisted. This might be a result of the 43-year-old legislation finally paying dividends, and because the Obama administration has put more resources into processing delistings to counter attacks from Republicans.

SPECIES RECOVERY SURGES UNDER OBAMA

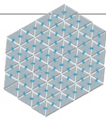
US President Barack Obama has overseen more species recoveries under the Endangered Species Act than all other presidents combined.



SOURCE: US DEPT OF INTERIOR/US FISH & WILDLIFE SERVICE

➲ NATURE.COM
For daily news updates see:
www.nature.com/news

NEWS IN FOCUS



DESMOND BOYLAN FOR *NATURE*

A fumigator sprays a Havana home with pesticide to kill mosquitoes. Eradication campaigns across Cuba have helped to ward off Zika until now.

EPIDEMIOLOGY

Cuba's epic battle with Zika

The country is one of the last in the Caribbean to get hit.

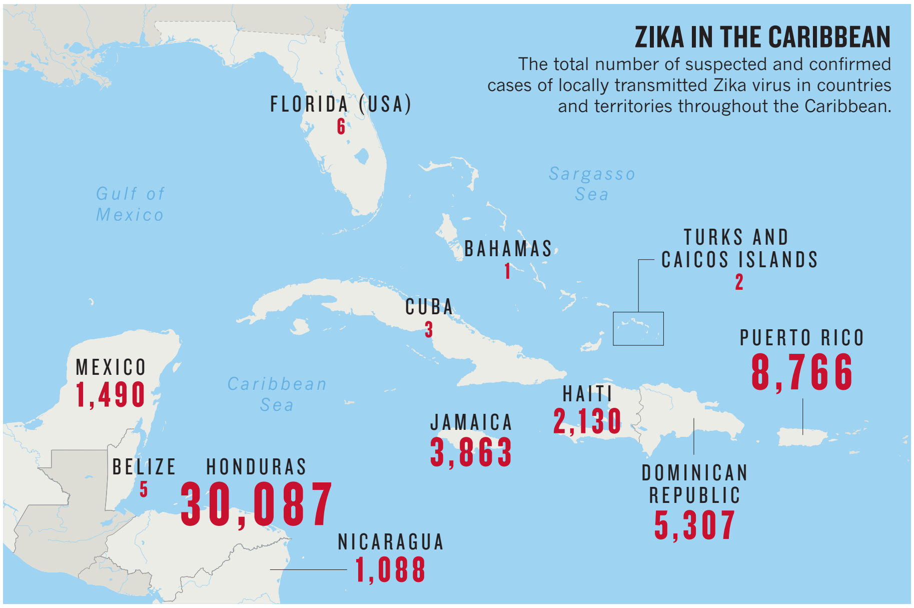
BY SARA REARDON

As soon as the rain stops, mosquitoes flood the guard house of an upscale tourist resort near Cuba's Bay of Pigs. Without hesitation, one of the guards reaches under his desk to pull out a device that looks like a very large hair dryer. "Mosquito gun," he says. He walks around, spraying a thick, white cloud of fumigant that engulfs the booth. Slowly, the mosquitoes disappear.

It's not uncommon to see clouds of pesticide wafting through Cuba's houses and neighbourhoods. It is largely because of such intensive measures by ordinary citizens that the country has been among the last in the Caribbean to succumb to local transmission of Zika. As of 11 August, Cuba has recorded three people who were infected by local mosquitoes rather than contracting the illness abroad, compared with 8,766 confirmed cases in nearby Puerto Rico (see 'Zika in the Caribbean').

Although scientists and public-health officials are disappointed that Zika has finally arrived in Cuba, they are not surprised. "It's not easy to avoid an introduction, because a lot of people are coming to Cuba from a lot of places," says Maria Guzmán, head of virology at the Pedro Kourí Tropical Medicine Institute in Havana. The country has recorded about 30 confirmed imported cases.

Zika is especially insidious because most people who have it show either no symptoms or only common ones such as fevers, which could be attributed to other illnesses. Yet, ▶



ZIKA IN THE CARIBBEAN

The total number of suspected and confirmed cases of locally transmitted Zika virus in countries and territories throughout the Caribbean.

SOURCE: PAHO/WHO

with the exception of one locally acquired case in March, Cuba mostly managed to keep Zika out until this month.

ON THE BALL

That success was the result of its excellent health-care system and an extensive surveillance programme for vector-borne diseases that the government set up 35 years ago, says Ileana Morales, director of science and technology at Cuba's public-health ministry.

In 1981, Cuba saw the first outbreak of haemorrhagic dengue fever in the Americas, with more than 344,000 infections. "We turned that epidemiological event into an opportunity," says Morales. The country sent medical workers to affected areas and began intensively spraying pesticides to eradicate the *Aedes aegypti* mosquito that carries the disease.

It also created a national reporting system, as well as a framework for cooperation between government agencies and public-education campaigns to encourage spraying and self-monitoring for mosquito bites, even among children. One of the most effective measures was a heavy fine for people found to have mosquitoes breeding on their property, says Duane Gubler, an infectious-disease researcher at Duke–NUS Medical School in Singapore. With all these measures in place, Cuba eliminated the dengue outbreak in four months.

Now, when another outbreak threatens, "it's no problem for us to reinforce our system" and intensify such efforts, says Morales.

In February, before any Zika cases had been detected in Cuba, the government dispatched 9,000 soldiers to spray homes and other buildings, while workers killed mosquito larvae in habitats such as waterways. Airport officials screened visitors arriving from Zika-infected countries and medical workers went from door to door looking for people with symptoms. The health-care system already conducts extensive prenatal examinations, so it is primed to detect Zika-caused birth defects such as microcephaly.

Cristian Morales, head of the Cuba office of the Pan American Health Organization (PAHO), says that it is probably unrealistic for other countries to simply copy Cuba's mosquito-control programmes. The country's health-care network is one of the best in the developing world, and the decades-long stability of its government has ensured policy continuity and enforcement of measures such as fines. He adds that the most important aspects of a response, for any country, include collaboration between government sectors and increased surveillance.

EVERYONE'S CHALLENGE

"Cuba probably does a better job of controlling mosquitoes than any other country in the Americas, but it hasn't been totally effective," says Gubler. This is partly due to dips in funding. A resurgence of dengue in 1997 was probably exacerbated by the fall of the Soviet Union, Cuba's major trading partner, which decimated the economy and weakened health funding.

Another disadvantage stems from the 56-year-old US trade embargo, which prevents Cuba from acquiring drugs and medical supplies that include components made in the United States. It must instead buy them from other countries, such as China, often at higher cost.

Yet success has come despite these issues. According to PAHO, health workers have intensified efforts to spray pesticides and eradicate standing water — where mosquitoes can breed — within 150 metres of the homes of each of the two most recent people to get Zika, in the southeastern province of Holguín. Workers are also searching houses for infected people and collecting mosquitoes for study. Guzmán adds that Cuban researchers have begun to plan work on a Zika vaccine.

She says that international cooperation will be important in helping Cuba and others to address Zika. "It's a problem of everybody. It's a new challenge for the world." ■

PHYSICS

Black–hole mimic triumphs

Result could be closest thing yet to observation of Hawking radiation.

BY DAVIDE CASTELVECCHI

Black holes are not actually black. Instead, these gravitational sinks are thought to emit radiation that causes them to shrink and eventually disappear. This phenomenon, one of the weirdest things about black holes, was predicted by Stephen Hawking more than 40 years ago, creating problems for theoretical physics that still convulse the field.

Now, after seven years of often solitary study, Jeff Steinhauer, an experimental physicist at the Technion-Israel Institute of Technology in Haifa, has created an artificial black hole that seems to emit such 'Hawking radiation' on its own, from quantum fluctuations that emerge from its experimental set-up.

It is nearly impossible to observe Hawking radiation in a real black hole, and previous artificial-black-hole experiments did not trace their radiation to spontaneous fluctuations. So the result, published on 15 August[1], could be the closest thing yet to an observation of Hawking radiation.

Steinhauer says that black-hole analogues might help to solve some of the dilemmas that the phenomenon poses for other theories, including one called the black-hole information paradox, and perhaps point the way to uniting quantum mechanics with a theory of gravity.

Other physicists are impressed, but they caution that the results are not clear-cut. And some doubt whether laboratory analogues can reveal much about real black holes. “This experiment, if all statements hold, is really amazing,” says Silke Weinfurtner, a theoretical and experimental physicist at the University of Nottingham, UK. “It doesn’t prove that Hawking radiation exists around astrophysical black holes.”

It was in the mid-1970s that Hawking, a theoretical physicist at the University of Cambridge, UK, discovered that the event horizon of a black hole — the surface from which nothing, including light, can escape — should have peculiar consequences for physics.

His starting point was that the randomness of quantum theory ruled out the existence of true nothingness. Even the emptiest region of space teems with fluctuations in energy fields, causing photon pairs to appear continuously, only to immediately destroy each other. But, just as Pinocchio turned from a puppet into a boy, these ‘virtual’ photons could become real particles if the event horizon separated them before they could annihilate each other. One photon would fall inside the event horizon and the other would escape into outer space.

This, Hawking showed, causes black holes both to radiate — albeit extremely feebly — and to ultimately shrink and vanish, because the particle that falls inside always has a ‘negative energy’ that depletes the black hole. Most controversially, Hawking also suggested that a black hole’s disappearance destroys all information about objects that have fallen into it, contradicting the accepted wisdom that the total amount of information in the Universe stays constant.

In the early 1980s, physicist Bill Unruh of the University of British Columbia in Vancouver, Canada, proposed a way to test some of Hawking’s predictions[2]. He imagined a medium that experienced accelerated motion, such as water approaching a waterfall. Like a swimmer reaching a point where he cannot swim away fast enough to escape the waterfall, sound waves that are past the point in the medium that surpasses the speed of sound would become unable to move against the flow. Unruh predicted that this point is equivalent to an event horizon — and that it should display a sonic form of Hawking radiation.

“For sure, this is a pioneering paper.”

Steinhauer implemented Unruh’s idea in a cloud of rubidium atoms that he cooled to a fraction of a degree above absolute zero. Contained in a cigar-shaped trap a few millimetres long, the atoms entered a quantum state called a Bose–Einstein condensate (BEC), in which the speed of sound was just half a millimetre per second. Steinhauer created an event horizon by accelerating the atoms until some were travelling at more than 1 mm s^{-1} — a supersonic speed for the condensate (see ‘Building a black hole’).

At its ultracold temperature, the BEC undergoes only weak quantum fluctuations that are similar to those in the vacuum of space. And these should produce packets of sound called phonons, just as the vacuum produces photons, Steinhauer says. The partners should separate from each other, with one partner on the supersonic side of the horizon and the other forming Hawking radiation.

On one side of his acoustical event horizon, where the atoms move at supersonic speeds, phonons became trapped. And when Steinhauer took pictures of the BEC, he found correlations between the densities of atoms that were an equal distance from the event horizon but on opposite sides. This demonstrates that pairs of phonons were entangled — a sign that they originated spontaneously from the same quantum fluctuation, he says, and that the BEC was producing Hawking radiation.

➲ NATURE.COM
Read more about the physicist who models black holes in sound:
go.nature.com/2atqxhy

By contrast, radiation that he observed in an earlier version of the set-up had to be triggered, rather than arising from the BEC itself[3], whereas a previous experiment in water waves led by Unruh and Weinfurtner did not attempt to show quantum effects[4].

Just as real black holes are not black, Steinhauer’s acoustical black holes are not completely quiet. Their sound, if it were audible, might resemble static noise.

“For sure, this is a pioneering paper,” says Ulf Leonhardt, a physicist at the Weizmann Institute of Science in Rehovot, Israel, who leads a different attempt to demonstrate the effect, using laser waves in an optical fibre. But he says that the evidence of entanglement seems incomplete, because Steinhauer demonstrated correlations only for phonons of relatively high energies, with lower-energy phonon pairs seemingly not correlated. He also says he’s not confident that the medium is a true BEC, which, he says, means that there could be other types of fluctuation that could mimic Hawking radiation.

Also unclear is what analogues can say about the mysteries surrounding true black holes. “I don’t believe it will illuminate the so-called information paradox,” says Leonard Susskind, a theoretical physicist at Stanford University in California. In contrast to the case of astrophysical black holes, there is no information loss in Steinhauer’s sonic black hole because the BEC does not evaporate.

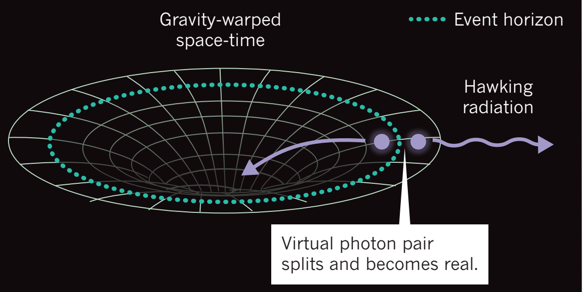
Still, if Steinhauer’s results were confirmed, it would be “a triumph for Hawking, perhaps in the same sense that the expected detection of the Higgs boson was a triumph for Higgs and company”, says Susskind. Few doubted that the particle existed, but its discovery in 2012 still earned Peter Higgs and another theorist, François Englert, who predicted it, a Nobel prize. ■

BUILDING A BLACK HOLE

A black hole’s event horizon — the point beyond which the gravitational pull is too strong even for light to escape — has been mimicked in the lab using a cloud of ultracold atoms. The artificial black hole seems to emit a form of Hawking radiation.

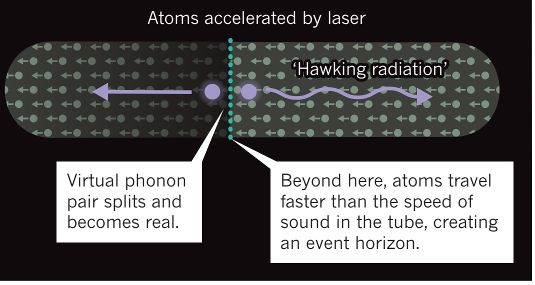
REAL BLACK HOLE

Quantum fluctuations in the vacuum of space produce virtual photons. Sometimes, one of a pair gets trapped behind the event horizon before the two destroy each other, forcing both to become real particles. The photon that escapes is emitted as Hawking radiation.


Gravity-warped space-time
Event horizon
Hawking radiation
Virtual photon pair splits and becomes real.

ARTIFICIAL BLACK HOLE

Ultracold atoms in a tube undergo quantum fluctuations that produce pairs of virtual particles — in this case, packets of sound called phonons. If one phonon falls in the supersonic region, it is trapped, leading to a sonic form of Hawking radiation.


Atoms accelerated by laser
‘Hawking radiation’
Virtual phonon pair splits and becomes real.
Beyond here, atoms travel faster than the speed of sound in the tube, creating an event horizon.

1. Steinhauer, J. *Nature Phys.* http://dx.doi.org/10.1038/nphys3863 (2016).
2. Unruh, W. G. *Phys. Rev. Lett.* **46,** 1351–1353 (1981).
3. Steinhauer, J. *Nature Phys.* **10,** 864–869 (2014).
4. Weinfurtner, S. *et al. Phys. Rev. Lett.* **106,** 021302 (2011).

RESEARCH INSTITUTES

Egypt science city in trouble

State support needed for project pioneered by Nobel laureate Ahmed Zewail.

BY PAKINAM AMER & MOHAMMED YAHIA

Questions are swirling over the future of Egypt's first science city, after the death of the Nobel laureate who made the project his legacy. The Zewail City of Science and Technology, a campus outside Cairo comprising a non-profit university and several research institutes, is named after the man who spearheaded it: Egyptian-born US chemist Ahmed Zewail, the first Arab to win a science Nobel prize.

But Zewail's death at the age of 70 on 2 August raises fresh doubts about the research hub's already precarious finances. The institute, which opened in 2011, had relied heavily on Zewail's star name and contacts to attract the support of scientific luminaries, as well as donations of 700 million Egyptian pounds (around US$80 million). It is now running out of that money, and, despite a loan of 1 billion Egyptian pounds from the ministry of defence, it has not raised enough cash to support a planned move to a new $450-million campus in 2019, says Sherif Fouad, a spokesman for the institute.

"Fundraising has always been a challenge, and I think it is likely to be affected by the loss of Dr Zewail in the short term," says Sherif El-Khamisy, a molecular biologist at the University of Sheffield, UK, who is also director of Zewail City's Center for Genomics. But he and others affiliated with the hub say they are hopeful that it will survive. In a speech on 6 August after Zewail's death, Egypt's president, Abdel Fattah el-Sisi, asked Egyptians to continue to donate to the city, but vowed that the nation's armed forces — whose engineers are building the new campus — would finish construction even if no more money comes through.

It is likely that Egypt's government will ultimately need to step in with support, says Salah Obayya, a physicist who is acting as chairman of Zewail City until a replacement for Zewail is elected. "The logistical support envisaged from the state is expected to override the initial fear or uncertainty," says El-Khamisy. ■

See go.nature.com/2bthapb for a longer version of this story.

George W. Bush had barriers erected along nearly 1,100 kilometres of frontier during his presidency.

CHARLES OMMANNEY/GETTY REPORTAGE

ECOLOGY

Trump's border-wall pledge raises hackles

Ecologists fear plan to seal off the United States from Mexico would put wildlife at risk.

BY BRIAN OWENS

With Republican presidential candidate Donald Trump talking about walling off the United States from Mexico, ecologists fear for the future of the delicate and surprisingly diverse ecosystems that span Mexico's border with the southwestern United States.

"The southwestern US and northwestern Mexico share their weather, rivers and wildlife," says Sergio Avila-Villegas, a conservation scientist from the Arizona-Sonora Desert Museum in Tucson. "The infrastructure on the border cuts through all that and divides a shared landscape in two."

Trump's policies tend to be short on detail, but he has talked about sealing off the entire 3,200-kilometre border with a wall that would be 10–20 metres high. "We will build a wall," Trump says in a video on his campaign website. "It will be a great wall. It will do what it is supposed to do: keep illegal immigrants out."

Constructing a wall "would be a huge loss", says Clinton Epps, a wildlife biologist at Oregon State University in Corvallis. "We know how important the natural movement of wildlife is for the persistence of many species."

Far from being a barren wasteland, the US–Mexico borderlands have some of the highest diversity of mammals, birds and plants in the continental United States and northern Mexico — including many threatened species.

A wall could divide species that make a home in both nations. Bighorn sheep, for example, live in small groups and rely on cross-border connections to survive, says Epps. Other species, such as jaguars, ocelots and bears, are concentrated in Mexico but have smaller, genetically linked US populations.

"Black bears were extirpated in West Texas, and it was a big deal when they re-established in the 1990s," Epps says. Breaking their links with Mexican bears could put the animals at risk again. And birds that rarely fly, such as roadrunners, or those that swoop low to the ground, such as pygmy owls, could also have trouble surmounting the wall.

Such a physical barrier would worsen the habitat disruption caused by noise, bright lights and traffic near the border. And a wall would cut across rivers and streams that cross the border, severing a vital link. "When water crosses the border, it unites ecosystems," says Avila-Villegas. "If we block the water, it affects nature on a much more fundamental level."

Trump is not the first US politician to hit upon the idea of sealing the southern border. In 2006, President George W. Bush authorized the construction of a 1,126-kilometre border wall, of which nearly 1,100 kilometres were completed. The existing barriers are a mixture of 6-metre-high steel walls, 'bollard fences' made of steel pipes set upright in the ground about 5 centimetres apart, and lower vehicle barriers that Avila-Villegas says resemble the tank traps set on the beaches of Normandy during the Second World War.

Few studies have explored these barriers' effects on animal populations, and there are not even any reliable baseline data on conditions before the barriers were built. Avila-Villegas has seen photos taken by border patrols of mountain lions running alongside the barriers or trying to climb over them, so he knows that the walls are causing the animals stress. But he has no real way of measuring it. A 2014 study found that the fencing in Arizona seemed to harm native wildlife, but had little impact on human movement (J. W. McCallum *et al. PLoS ONE* **9,** e93679; 2014).

In 2009, Epps published a paper setting out some of the potential threats to animal populations posed by Bush's wall, but he lacked the money to follow up with field studies (A. D. Flesch *et al. Conserv. Biol.* **24,** 171–181; 2009). Now he is not sure such research would be possible, even with sufficient funds. "The border is not a friendly place any more," Epps says. "I would be hesitant to send a grad student there."

Avila-Villegas has first-hand experience of the difficulties that researchers face there. Ten years ago, he tried to collect some baseline data before Bush's barriers were built, but gave up for his own safety. "It's easy to ask why the research hasn't been done, but that ignores the fact that the border is a war zone," he says. "I had to stop my field work because of law enforcement and the Minutemen" — groups of armed private citizens who have taken it upon themselves to 'defend' the border against illegal crossings.

And it has not got any easier. "Every time I — a Hispanic male with dark skin and long hair — am in the field, I get patrols, helicopters and ATVs [all-terrain vehicles] coming to check on what I'm doing," Avila-Villegas says. He spends much of his time trying to promote conservation issues that affect Mexico and the United States by forging links between researchers and policymakers in both countries. But his dedication to an open border has also prompted him to take a more personal stand. After a dozen years in the United States, Avila-Villegas has finally applied for citizenship — so that, come November, he can vote against Trump and his wall. ■

PHYSICS

Neutrino clue to Universe riddle

Hint that elusive particles behave differently in matter and antimatter forms might explain matter's predominance.

BY ELIZABETH GIBNEY

It is one of physics' greatest mysteries: why the Universe is filled with matter, rather than antimatter. An experiment in Japan now hints at a possible explanation: subatomic particles called neutrinos might behave differently in their matter and antimatter forms.

The disparity, announced at the International Conference on High Energy Physics (ICHEP) in Chicago, Illinois, on 6 August, may yet turn out not to be real: more data will need to be gathered to be sure. "You would probably bet that this difference exists in neutrinos, but it would be premature to state that we can see it," says André de Gouvêa, a theoretical physicist at Northwestern University in Evanston, Illinois.

Even so, the announcement is likely to increase excitement over studies of neutrinos, the abundant but elusive particles that seem increasingly key to solving all kinds of puzzles in physics.

"For the timescales of particle physics, this is changing really, really quickly."

In the 1990s, neutrinos were found to defy the predictions of physics' standard model — a successful, but incomplete, description of nature — by virtue of possessing mass, rather than being entirely massless (Y. Fukuda *et al. Phys. Rev. Lett.* **81,** 1562; 1998). Since then, neutrino experiments have sprouted up around the world, and researchers are realizing that they should look to these particles for new explanations in physics, says Keith Matera, a physicist on a US-based neutrino experiment called NOvA at the Fermi National Accelerator Laboratory (Fermilab) in Batavia, Illinois. "They are the crack in the standard model," he says.

If matter and antimatter were produced in equal quantities after the Big Bang, they would have annihilated each other, leaving nothing but radiation. Physicists have observed differences in the behaviour of some matter particles and antimatter particles, such as kaons and B mesons — but not enough to explain the dominance of matter in the Universe.

AN ODD ABUNDANCE

One answer might be that super-heavy particles decayed in the early Universe in an asymmetrical fashion and produced more matter than antimatter. Some physicists think that a heavyweight relative of the neutrino could be the culprit. Under this theory, if neutrinos and antineutrinos behave differently today, then a similar imbalance in their ancient counterparts could explain the overabundance of matter.

To test this, researchers on the Tokai to Kamioka (T2K) experiment in Japan looked for differences in the way that matter and antimatter neutrinos oscillate between three types, or 'flavours', as they travel. They shot beams of neutrinos of one flavour — muon neutrinos — from the Japan Proton Accelerator Research Complex in the seaside village of Tokaimura to the Super-Kamiokande detector, an underground steel tank more than 295 kilometres away and filled with 50,000 tonnes of water. The team counted how many electron neutrinos appeared — a sign that the muon neutrinos had morphed into a different flavour along the journey. They then repeated the experiment with a beam of muon antineutrinos.

▶

MORE ONLINE

Q&A

MINISTERIO DE CIENCIA, TECNOLOGÍA E INNOVACIÓN PRODUCTIVA

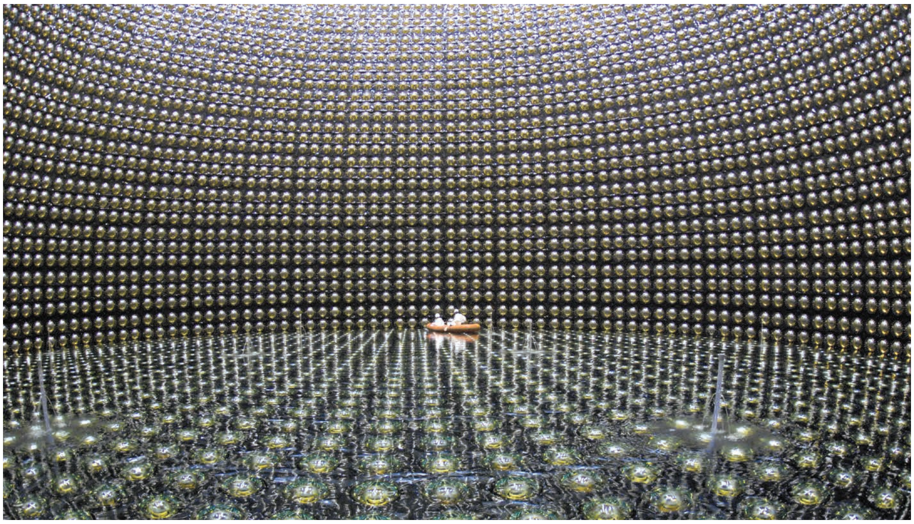
The man behind Argentina's science growth go.nature.com/2buhxxm

MORE NEWS

- Scientists probe lightning hotspot to predict strikes go.nature.com/2bv1vcd
- Why the US decision to expand marijuana supply for research matters go.nature.com/2aydj72
- Near-blind shark is world's longest-lived vertebrate go.nature.com/2awg5an

NATURE PODCAST

How fins became limbs; giant gene database cracks clinical cases; and making better opioids nature.com/nature/podcast



Japan's Super-Kamiokande detector near Hida is analysing matter and antimatter neutrinos.

KAMIOKA OBSERVATORY, ICRR (INST. FOR COSMIC RAY RESEARCH), UNIV. TOKYO

The two beams behaved slightly differently, said Konosuke Iwamoto, a physicist at the University of Rochester, New York, during his presentation at ICHEP.

The team expected that if there were no difference between matter and antimatter, their detector would have, after almost 6 years of experiments, seen 24 electron neutrinos and — because antimatter is harder to produce and detect — 7 electron antineutrinos. Instead, they saw 32 neutrinos and 4 antineutrinos arrive in their detector. "Without getting into complicated mathematics, this suggests that matter and antimatter do not oscillate in the same way," says Chang Kee Jung, a physicist at Stony Brook University in New York and a member of the T2K experiment.

Preliminary findings from the T2K and NOvA experiments had hinted at the same idea. But the observations so far could be chance fluctuations; there is a 1 in 20 chance (or in statistical terms, about 2 sigma) of seeing these results if neutrinos and antineutrinos behave identically, says Jung. By the end of its current run in 2021, the T2K experiment should have five times more data than it has today. But the team will need 13 times more data to push statistical confidence in the finding to 3 sigma, a statistical threshold beyond which most physicists would accept the data as reasonable — but not completely convincing — evidence of the asymmetry.

The T2K team has proposed extending its experiment to 2025 to gather the necessary data. But it is trying to speed up data-gathering by combining results with those from NOvA, which sends a neutrino beam 810 kilometres from Fermilab to a mine in northern Minnesota. NOvA has been shooting neutrino beams; it will switch to antineutrino beams in 2017. The two groups have agreed to produce a joint analysis and could together reach 3 sigma by around 2020, says Jung. Reaching the statistical certainty needed for a formal discovery, 5 sigma, might require a new generation of neutrino experiments, which are already being planned around the world.

Physicists are racking up discoveries about neutrinos on an almost annual basis, says de Gouvêa. "For the timescales of particle physics, this is changing really, really quickly." ■

BOB DAEMMRICH/POLARIS/EYEVINE

THE PLASTIC OCEAN

SCIENTISTS KNOW THAT THERE IS A COLOSSAL AMOUNT OF PLASTIC IN THE OCEANS. BUT THEY DON'T KNOW WHERE IT ALL IS, WHAT IT LOOKS LIKE OR WHAT DAMAGE IT DOES.

BY DANIEL CRESSEY

Kamilo beach, on the tip of Hawaii's Big Island, is a remote tropical shore. It has white sand, powerful waves and cannot be reached by road. It has, in fact, much that an idyllic tropical beach should have. But there is one inescapable issue: it is regularly carpeted with plastic.

Bottles, fishing nets, ropes, shoes and toothbrushes are among the tons of waste washed up here, thanks to a combination of ocean currents and local eddies. A study in 2011 reported that the top sand layer could be up to 30% plastic by weight[1]. It has been called the dirtiest beach in the world, and is a startling and visible demonstration of how much plastic detritus humanity has dumped into the world's oceans.

From Arctic to Antarctic, from surface to sediment, in every marine environment where scientists have looked, they have found plastic. Other human-generated debris rots or rusts away, but plastics can persist for years, killing animals, polluting the environment and blighting coastlines. By some estimates, plastics comprise 50–80% of the litter in the oceans. "There are places where you don't find plastic," says Kara Lavender Law, an oceanographer at the Sea Education Association in Woods Hole, Massachusetts. "But in terms of the different marine reservoirs, we've found plastic in all of them. We know it's pervasive."

Newspapers tell stories of the 'Great Pacific garbage patch', a region of the central Pacific where plastic particles accumulate, and volunteers participate in beach clean-ups across the globe. But in many ways, research lags behind public concern. Scientists are still struggling to answer the most basic questions: how much plastic is in the oceans, where, in what form and what harm it's doing. That's because science at sea is hard, expensive and time-consuming. It is difficult to comprehensively survey vast oceans for small — sometimes microscopic — plastic fragments, and few researchers have made this their line of work.

But now interest is picking up. "There have been more publications in the last four years than the previous four decades," says Marcus Eriksen, director of research and co-founder of the 5 Gyres Institute in Santa Monica, California, which works to fight plastic pollution. Scientists and environmentalists know that there is a lot to do. Last May, the United Nations Environment Programme (UNEP) passed a resolution at its Nairobi meeting, stating that "the presence of plastic litter and microplastics in the marine environment is a rapidly increasing serious issue of global concern that needs an urgent global response".

WHERE DOES IT COME FROM?

In 2014, a team at the US marine park Papahānaumokuākea, off the northwest coast of Hawaii, removed a fishing net from the reserve that weighed 11.5 tonnes — roughly equivalent to a London bus. Nets and other fishing equipment that have been lost or discarded at sea are thought to make up a large fraction of marine plastic. An estimate[2] from UNEP suggests that this 'ghost' fishing gear makes up 10% of all marine litter, or around 640,000 tonnes.

There is much more than that. Global production of plastics rises every year — it is now up to around 300 million tonnes — and much of it eventually ends up in the ocean. Plastic litter is left on beaches, and plastic bags blow into the sea. The vast quantities of plastics dumped as landfill can, if sites are not properly managed, easily wash or blow away. Some sources are less obvious: as tyres wear down, they leave tiny fragments on roads that leach into drains and on into the ocean.

In a 2014 paper, Eriksen and his team analysed data on the items found in a series of expeditions across the world's oceans and estimated that 87% by weight of floating plastic was greater than 4.75 millimetres in size[3]. The list included buoys, lines, nets, buckets, bottles and bags (see 'A sea of plastic'). But when the pieces were counted instead of weighed, large plastics made up just 7% of the total. Many plastic items break down under the onslaught of sunlight and waves until they eventually reach microscopic sizes, and other plastics are small from the start, such as the 'microbeads' that are added to face scrubs and other cosmetic products, and that go down the drain.

Concern about these microplastics has been growing ever since 2004, when Richard Thompson, who researches ocean plastic at Plymouth University in the United Kingdom, coined the term. (It is now often used to refer to pieces less than 5 millimetres across.) His team found microplastics in most of the samples it took from 18 British beaches, as well as in plankton samples collected from the North Sea as far back as the 1960s[4]. Since then, the number of papers using the term has rocketed, and researchers are attempting to answer questions ranging from how toxic the materials are, to how they are distributed around the world.

HOW MUCH IS OUT THERE?

If surveying the ocean for plastic is expensive and difficult at the surface, it's even harder below it: researchers lack samples from enormous areas of the deep sea that have never been explored. And even if they could survey all these regions, the concentration is typically so dilute that they would have to test huge volumes of water to get reliable results. Instead, they are forced to estimate and extrapolate.

In a paper published last year, a team led by Jenna Jambeck, who researches waste management at the University of Georgia in Athens, estimated how much waste coastal countries and territories generate, and how much of that could be plastic that ends up in the ocean[5]. The group reached a figure of 4.8 million to 12.7 million tonnes every year — very roughly equivalent to 500 billion plastic drinks bottles. But her estimate excluded the plastic that gets lost or dumped at sea, and all the plastic that is already there.

To get a handle on this, some researchers have gone trawling, using fine-meshed nets to see what plastic they can catch. Last year, oceanographer Erik van Sebille of Imperial College London and his colleagues published one of the largest collections of such data[6]. They combined information from 11,854 individual trawls, from every ocean except the Arctic, to produce a 'global inventory' of small plastic pieces floating at or near the surface.

They estimated that, in 2014, there were between 15 trillion and 51 trillion pieces of microplastic floating in the oceans, with a total weight of 93,000 to 236,000 tonnes. But these numbers present scientists with a problem. This estimate of total surface plastic is just a small fraction of what Jambeck estimated entered the ocean every year. So where is all the rest? "That's the big question," says Jambeck. "That's a tough one."

Researchers are trying to find answers. Jambeck is now working with a mobile-phone app called the Marine Debris Tracker, which offers a way to crowdsource vast amounts of data as users send in information about rubbish they encounter. She is also working on a project for UNEP to build a global database of marine-litter projects.

WHERE IS IT?

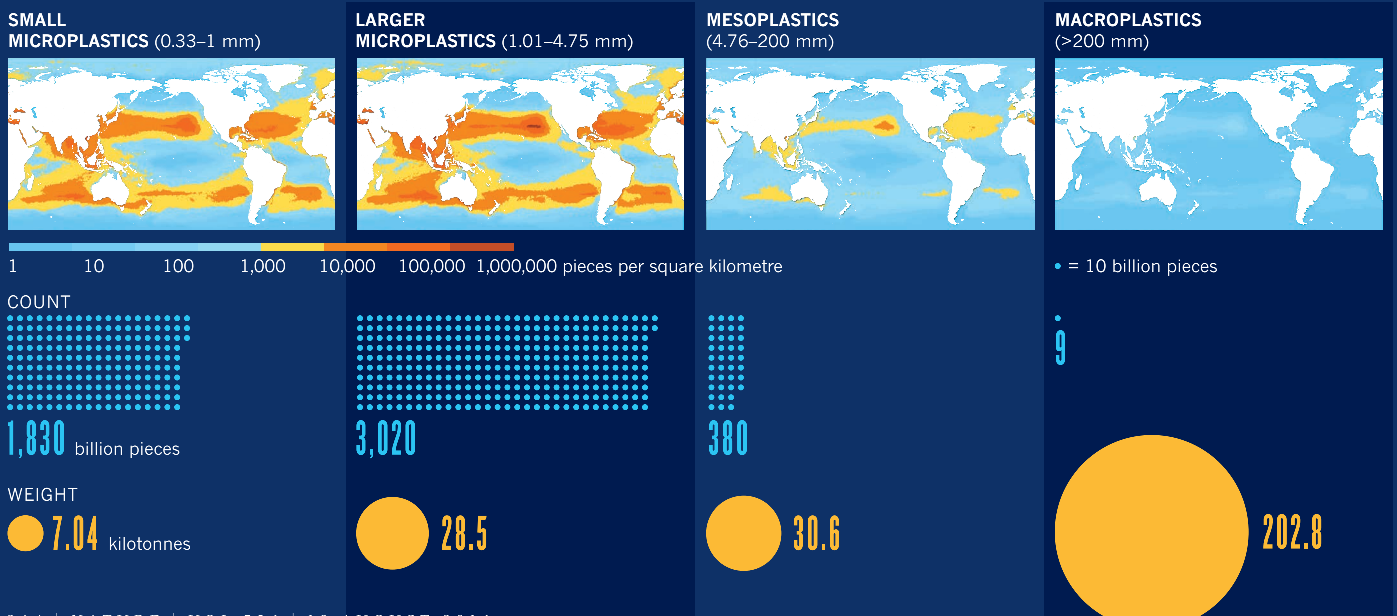
The mismatch between the estimated amount of plastic entering the oceans and the amount actually observed has come to be known as the 'missing plastic' problem. Adding to the puzzle, data from some locations do not show a clear increase in plastic concentrations over recent years, even though global production of the materials is soaring.

Public attention has focused on the Great Pacific garbage patch, where plastics collect thanks to an ocean current called a gyre. The name is something of a misnomer — visitors to the patch would not find piles of seaborne rubbish. A study from 2001 reported 334,271 pieces of plastic per square kilometre in the gyre[7]. This is the largest tally recorded in the Pacific Ocean, but still works out as roughly one small fragment for every three square metres.

Modelling by van Sebille and his colleagues suggest that concentrations could be several orders of magnitude higher in the Pacific garbage patch, and an equivalent zone in the North Atlantic, than elsewhere.

A SEA OF PLASTIC

A 2014 study[3] estimated that more than 5 trillion plastic pieces, weighing more than 250,000 tonnes, float on the surface of the world's oceans. Small pieces make up the majority by count, but large items account for the greatest weight. Currents cause plastics to accumulate in the North Pacific and North Atlantic 'garbage patches'.



But the plastic here is accounted for in surveys, whereas the missing plastic is, by definition, missing and therefore somewhere else.

Some of it is probably on the sea floor. Certain types of plastic sink, and even ones that start out floating can eventually become covered with marine organisms and be pulled down. Work from Thompson has shown microplastics in deep-ocean sediment — an under-studied zone that could be hiding some of the missing millions of tonnes[8]. Remotely operated vehicles also regularly find large plastic items among the litter that has sunk into the deepest ocean trenches.

A substantial portion of ocean plastic may simply end up on shorelines, and other plastic 'sinks' are uncovered all the time. In 2014, Thompson co-authored a paper showing that microplastics had accumulated in Arctic sea ice at concentrations several orders of magnitude greater than that found even in highly contaminated surface waters[9]. "We have a lot of educated guesses" about where the missing plastic is, says Law. "In my mind, we don't have the answer to that."

Thompson and others are now looking beyond microplastics to nanoplastics — ones less than 100 nanometres in size. "Nano-sized particles of plastic are being manufactured," says Thompson. "So it's highly likely that some will escape into the environment. There's also the fragmentation of larger items." But nanoplastics are proving hard to study. Researchers commonly use a type of spectroscopy to confirm whether fragments recovered from the sea are made of plastic, but the method does not work well on pieces below about 10 micrometres, Thompson says. He hopes to learn more as part of a UK-government-funded project called RealRiskNano, which will look at sources and pathways to the environment for these tiny fragments. "It wouldn't surprise me to find they do exist. But at the moment it's below the level of detection from an environmental sample."

WHAT HARM DOES IT DO?

Researchers know that marine plastic can harm animals. Ghost fishing gear has trapped and killed hundreds of animal species, from turtles to seals to birds. Many organisms also swallow pieces of plastic, which can accumulate in their digestive system. According to one often-quoted figure, around 90% of seabirds called fulmars washed ashore dead in the North Sea had plastic in their guts. What's less clear is whether this pollution has major impacts on populations.

Lab studies have demonstrated the toxicity of microplastics, but these often use concentrations that are much higher than those found in the oceans. In February this year, though, Arnaud Huvet, who studies invertebrates at France's national marine research agency (Ifremer) in Plouzané, published work in which he exposed Pacific oysters to microplastics at concentrations similar to those found in the sediment where the creatures live. Animals in the plastic-laced water had poorer-quality eggs and sperm and produced 41% fewer larvae than did those in a control group[10]. It was one of the first studies to show a direct link between plastic and fertility problems. "That made an impact," van Sebille says.

So did a study in June from fish ecologists Oona Lönnstedt and Peter Eklöv, in which they exposed perch larvae to 'environmentally relevant' concentrations of microplastics. The larvae ate the plastics — they even seemed to prefer them to actual food — which made them grow more slowly and fail to respond to the odour of predators. After 24 hours in a tank with a predator, 34% of plastic-dosed larvae survived, compared with 46% of those raised in clean water[11].

Lönnstedt, at Uppsala University in Sweden, was disturbed by photos of the transparent larvae clearly showing the small plastic spheres in their guts. "It's awful, so of course I feel strongly about it," she says. "People who say plastics won't be an issue in the oceans need to take a look at the evidence again."

But some scientists question the implications of the work. Alastair Grant, an ecologist at the University of East Anglia in Norwich, UK, says that the levels of plastic that gave adverse effects in Lönnstedt's paper — 10–80 particles per litre — are still orders of magnitude higher than the vast majority of field measurements. Most reports are less than 1 particle per litre, he says. "The evidence I can see at the moment suggests microplastics are probably within safe environmental limits in most places."

WHAT SHOULD WE DO?

Despite the lack of comprehensive data about ocean plastics, there is a broad consensus among researchers that humanity should not wait for more evidence before taking action. Then the question becomes, how?

One controversial project has been devised by The Ocean Cleanup, a non-profit group that by 2020 hopes to deploy a 100-kilometre-long floating barrier in the Great Pacific garbage patch. The group claims that the barrier will remove half of the surface plastic there.

But the project has met with scepticism from researchers. They say that plastic in the gyre is so dilute that it will be tough to scoop up, and they worry that the barrier will disturb fish populations and plankton. Boyan Slat, chief executive of The Ocean Cleanup, welcomes the criticism, but says that the barrier project is still in an early phase, with a prototype currently deployed off the Dutch coast. "We're using this test as a platform to investigate whether there's any negative consequences. The only way to find out is to go out and do it," he says.

> "WE'VE GOT TO STOP IT IN THE TREATMENT PLANTS. IN THE LANDFILLS. THAT IS THE POINT TO INTERVENE."

In a paper published earlier this year[12], van Sebille and his colleague Peter Sherman showed that it would be much more effective to place clean-up equipment near the coasts of China and Indonesia, where much of the plastic pollution originates. "The closer to the plastic economy loop you intervene the better it is," van Sebille says. "We've got to stop it in the treatment plants, in the landfills. That is the point to intervene." Eriksen likens the situation to addressing air pollution, where people have long realized that filtering the air is not a long-term solution. Filtering the oceans seems similarly implausible, he says. "What we've seen worldwide is you go to the source." That means reducing the use of plastic, improving waste management and recycling the materials to stop them from reaching the water at all.

That's a lot to ask, considering how ubiquitous plastics are. But some scientists allow themselves to imagine a world where plastics have been brought under control. According to research by Law and Jan van Franeker, some types of floating plastic might disappear in just a few years[13]. Perhaps even Kamilo beach would eventually return to its unpolluted form.

But plastic will have left its mark, as layers of tiny particles embedded in sediment on the ocean floor. Over time, this plastic will become cemented into Earth — a legacy of the plastic era. "There will be this layer of rock around the world that is going to be plastic," Eriksen says. ■ SEE NEWS FEATURE P.266, AND BOOKS AND ARTS P.272

Daniel Cressey *is a senior reporter for* Nature *in London.*

1. Carson, H. S., Colbert, S. L., Kaylor, M. J. & McDermid, K. J. *Mar. Pollut. Bull.* **62,** 1708–1713 (2011).
2. Macfadyen, G., Huntington, T. & Cappell, R. *Abandoned, Lost or Otherwise Discarded Fishing Gear* (UNEP, 2009).
3. Eriksen, M. *et al. PLoS ONE* **9,** e111913 (2014).
4. Thompson, R. C. *et al. Science* **304,** 838 (2004).
5. Jambeck, J. R. *et al. Science* **347,** 768–771 (2015).
6. van Sebille, E. *et al. Environ. Res. Lett.* **10,** 124006 (2015).
7. Moore, C. J., Moore, S. L., Leecaster, M. K. & Weisberg, S. B. *Mar. Pollut. Bull.* **42,** 1297–1300 (2001).
8. Woodall, L. C. *et al. R. Soc. Open Sci.* **1,** 140317 (2014).
9. Obbard, R. W., Sadri, S., Wong, Y. Q., Khitun, A. A., Baker, I. & Thompson, R. C. *Earth's Future* **2,** 315–320 (2014).
10. Sussarellu, R. *et al. Proc. Natl Acad. Sci. USA* **113,** 2430–2435 (2016).
11. Lönnstedt, O. M. & Eklöv, P. *Science* **352,** 1213–1216 (2016).
12. Sherman, P. & van Sebille, E. *Environ. Res. Lett.* **11,** 041001 (2016).
13. van Franeker, J. A. & Law, K. L. *Environ. Pollut.* **203,** 89–96 (2015).

NIK SPENCER/NATURE

Polymers have infiltrated almost every aspect of modern life. Now they are being stretched to their limits.

BY MARK PEPLOW

Hermann Staudinger was a pacifist, but this was one fight he was determined to win. In 1920, the German chemist proposed that polymers — a broad class of compounds that included rubber and cellulose — were made of long chains of identical small molecules linked by strong chemical bonds[1]. Most of his colleagues thought this was arrant nonsense, and argued that polymers were merely looser aggregations of small molecules. Staudinger refused to back down, sparking feuds that spanned a decade.

Eventually, laboratory data proved that he was right. He won the 1953 Nobel Prize in Chemistry for his work, and synthetic polymers are now ubiquitous: last year, the world produced about 300 million tonnes of them. The molecular chains that Staudinger hypothesized have entered almost every aspect of modern life, from clothes, paint and packaging to drug delivery, 3D printing and self-healing materials. Polymer-based composites even make up half the weight of Boeing's most recent passenger aeroplane, the 787 Dreamliner.

So where will polymers go next? Some answers will come this week, when a once-per-decade workshop organized by the US National Science Foundation attempts to survey which new areas are emerging.

"The general trend — still continuing — is the expansion of polymers into applications that have not been traditionally theirs," says Tim Lodge, a polymer chemist at the University of Minnesota in Minneapolis and editor of the journal *Macromolecules*. That expansion has been driven by advances in every aspect of polymer science, he says. Researchers have developed new methods to synthesize and analyse molecules, improved theoretical models and created mimics of polymers found in nature. At the same time, says Lodge, attitudes to the science have changed. No longer do universities dismiss polymer science as too dirty, practical and industrial for academia. "Just about every chemistry department has someone doing polymer stuff now," he says, and frontier work on polymers is increasingly interdisciplinary.

It will need to be. Researchers have a growing toolbox of techniques with which to craft the chemical architecture of polymer strands, but they are often unable to predict whether the resulting compound will have the particular properties needed for, say, a membrane or a drug-delivery system. Meeting that challenge will demand a much deeper understanding of how the chemical structure of a polymer determines its physical properties, at every scale from nanometres to metres.

POLYMERS FOREVER

Polymers are everywhere — and therein lies the problem. "Most polymers we use in everyday life are from petroleum-based products, and although they're durable in use, they're also durable in waste," says Marc Hillmyer, director of the Center for Sustainable Polymers (CSP) at the University of Minnesota. An estimated 86% of all plastic packaging is used only once before it is discarded[2], producing a stream of waste that persists in waterways and landfill, releases pollutants and harms wildlife (see page 263).

That is why the past decade has seen an explosion of interest in polymers that are made from renewable resources and biodegrade easily and harmlessly. Polymers based on natural starch are already on the market; so too is synthetic polylactide (PLA), which is made from lactide or lactic acid derived from biological sources, and which is found in products from tea bags to medical implants.

But sustainable polymers still make up less than 10% of the total plastics market, says Hillmyer. One hurdle is that they cost too much. Another is that the monomer building blocks of natural polymers tend to contain more oxygen atoms than are found in the fossil hydrocarbons of petroleum. This affects the polymers' properties — stiffening the materials, for example — which can make it difficult for them to directly replace cheap and flexible plastics such as polyethylene and polypropylene. Turning natural polymers into exact molecular matches for conventional ones takes some sophisticated chemistry.

One alternative approach is to beef up sustainable polymers such as PLA by blending them with conventional polymers. This route typically has downsides, such as rendering some plastics less transparent. But CSP researchers have got around that problem by adding just 5% by weight of a low-cost, petroleum-derived polymer that contains some sections that are hydrophobic — water-insoluble — and others that are hydrophilic, or water-soluble[3]. These additives cluster together to create spherical structures, which render PLA substantially tougher without reducing its transparency.

Hillmyer's team has also made[4] a partially recyclable form of polyurethane foam, which is found in a host of products, including insulation, seat cushions and gaskets. The recipe for this polyurethane includes a low-cost polymer called poly(β-methyl-δ-valerolactone) (PMVL), based on monomers made by modified bacteria. Heating the foam to above 200 °C breaks down the polyurethane so that the monomers can be extracted and used again.

It remains to be seen whether these sustainable polymers can be commercialized. "Often the biggest challenge is to do it at scale, which requires favourable economics," says Hillmyer. He thinks the field needs to establish general design rules that predict how a monomer's chemical structure affects the rate, temperature and yield of polymerization reactions, and how the resulting polymers will interact with other materials. His team has developed such guidelines for PMVL's constituents[5], and last year formed a spin-off company at the CSP called Valerian Materials to exploit these principles.

Some researchers are pursuing another trick: rather than stringing together bioderived monomers, they are learning to use natural polymers directly. Cellulose, for example, consists of glucose molecules strung together into chains, which in turn line up to form strong fibres, or fibrils, that make up the stiff cell walls of plants. In many places, the cellulose chains form crystalline chunks that are up to 20 nanometres wide and hundreds of nanometres long, and that can be chemically extracted from cellulose pulp. Proponents say that these crystals could be used for applications such as strengthening composites, forming insulating foams, delivering drugs and providing a scaffold for tissue repair[6].

> **"Just about every chemistry department has someone doing polymer stuff now."**

Cellulose nanocrystals and longer nanofibrils are now produced on a commercial scale, but the commercial applications do not yet go much beyond stiffening paper or thickening fluids. Christoph Weder, director of the Adolphe Merkle Institute for nanoscience at the University of Fribourg in Switzerland, says that it will take a lot more work to reduce costs and demonstrate unique advantages for sustainable polymers. "We really need a road map for biobased polymers," he says.

SKIN IN THE GAME

In a mixed-up world, polymers can restore some order. Polymer membranes already serve as molecular sieves for separating gases, desalinating seawater and keeping molecules apart inside fuel cells. But they could have a much bigger impact in the future, says Lodge. "There are so many problems that could be solved by better membranes."

Separating mixtures with membranes takes a lot less energy than does distillation, in which a liquid is heated to evaporate its components at different temperatures. It also requires much less space than using scrubbers, devices in which pollutants are trapped by chemical reactions. Membranes made from polymers are not only cheap to make at large scale, but can cover large areas without acquiring structural defects that let the wrong molecules pass through.

Gas-separation membranes are already used industrially to tease hydrogen and carbon dioxide from natural gas. But improved membranes could tackle harder tasks, such as distinguishing between the very similar hydrocarbons propane and propene. Tougher, chemically robust membranes could operate at higher temperatures to remove carbon dioxide from hot flue gases.

Membrane chemist Benny Freeman of the University of Texas at Austin is hoping to improve the treatment of waste water from gas fracking operations, in which water is forced into rock to split it open and release natural gas. After use, the water is so dirty that standard filtration membranes quickly get clogged, so the water must be put under high pressure to push it through, and the membranes must be cleaned with chemicals that shorten their lifespan. But Freeman has found a way to sidestep that problem by giving the membranes a gossamer-thin coating of polydopamine, which mimics the waterproof glue used by mussels to cling onto rocks. Piloted at a fracking water-treatment facility near Fort Worth, Texas, the polydopamine coating halved the pressure needed to push water through the membrane, which could result in smaller, more efficient treatment systems[7]. The team has already used these membranes to build units for the US Navy, so that ships can purify oily bilge water before dumping it.

In December 2015, the US presidential administration launched a 'moonshot for water' to boost water sustainability, and as part of that effort the US Department of Energy plans to establish a desalination-research hub in 2017. Polymer membranes will have a big role in that effort, says Freeman. "We're slated to see a huge increase in efforts to expand the use of polymers in that space."

To design better desalination membranes, researchers will need to be able to predict how factors such as the distribution of charged chemical groups in a polymer affects its permeability to ions. Earlier this year, Freeman and his colleagues published[8] what he believes is the first model to do just that, which could enable chemists to build particular properties into a membrane by tailoring its chemical substituents and cross-linking the molecules. "I'm on a mission to get people to ask these kinds of questions about structure–property relations, which could really guide synthesis," he says.

The ultimate separation membrane could be just one molecule thick. These 2D polymers are surfing the wave of enthusiasm for single-layer materials that followed the isolation of graphene just over a decade ago.

The flat polymers are not just very thin films of ordinary, linear polymers. Instead, they have an intrinsically 2D chemical structure that looks like a fishing net, with a regular, repeating mesh full of molecule-size openings. They can also carry a wide variety of chemical decorations on their surfaces, so that each opening can be precisely engineered to allow certain molecules through and bar others.

But creating 2D polymers is tough. If just one of the holes in the growing mesh closes up in the wrong way, the membrane could buckle into a 3D mess. Polymer chemist Dieter Schlüter of the Swiss Federal Institute of Technology in Zurich worked on this problem for more than a decade before achieving success in 2014.

His approach relies on coaxing carefully designed monomers to form a crystal. A blast of blue light then triggers a chemical reaction between monomers in the same plane, creating a new crystal made up of stacked polymer layers. These can be peeled off to give individual 2D sheets just one monomer thick (see 'Chemical peel').

Using the same approach, Schlüter and Benjamin King, head of the chemistry department at the University of Nevada, Reno, have independently produced different types of 2D polymer[9,10]. Now collaborators, the two researchers hope that they will soon be able to make these sheets in kilogram batches, easily enough to distribute samples to research groups around the world.

Schlüter admits that he has faced scepticism about whether 2D polymers will flourish. "But that's healthy," he says. "And I'm very stubborn — I will not give up, I'm convinced of the great potential this development has."

BOUTIQUE POLYMERS

Widely used polymers such as polystyrene and polyethylene are spectacularly boring in one sense: they repeat the same monomer over and over again. Their one-note tune is especially monotonous when compared with the quadraphonic symphony of DNA, which encodes an entire genome with 4 monomers; or the baroque masterpiece of a protein, drawing from 23 amino acids to build a complex 3D structure.

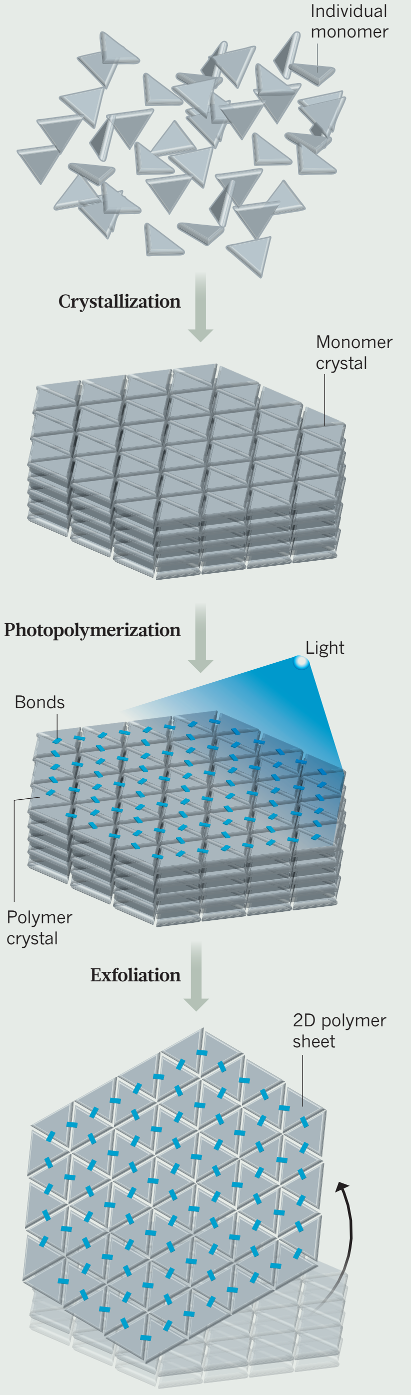
One of the most challenging frontiers of polymer research is to tailor synthetic polymers with the same precision, so that chemists can fine-tune the electronic and physical properties of their products. "It's become very fashionable in the past five years," says Jean-François Lutz, a macromolecular chemist at the University of Strasbourg in France. Sequence-controlled polymers would contain monomers in a predetermined order, forming strands of a very specific length.

Last year, a team led by Jeremiah Johnson, a chemist at the Massachusetts Institute of Technology in Cambridge, showed[11] that they could achieve that kind of control through iterative exponential growth — first uniting two different monomers to make a dimer, then connecting two dimers to make a tetramer, and so on. Modifying each monomer's chemical side-chains between cycles adds complexity, and a semi-automated system can make the process less laborious[12].

Johnson is now studying how his sequence-controlled polymers might be used in drug delivery. A dozen drugs approved by the US Food and Drug Administration use a polymer called polyethylene glycol to shield them from the body's immune system, improve their solubility or prolong their time in the body. Johnson says that a sequence-controlled polymer could provide a more predictable biological effect, because every strand would be the same length and shape, and its chemistry could be carefully designed to assist its drug cargo in the most useful way.

CHEMICAL PEEL

To make 2D polymer membranes, chemists first coax the monomers to form a 3D crystal, then shine a blue light to link monomers lying in the same plane. This allows them to peel off one sheet at a time.



Sequence-controlled polymers could also store data in a more compact and inexpensive form than can conventional semiconductor technology, with each monomer representing a single bit of information. Last year, Lutz demonstrated[13] a key step towards that goal. He used two types of monomer to represent digital 1s or 0s, and a third to act as a spacer between them. The monomers contained chemical groups that allowed them to connect only to the growing polymer, rather than reacting with each other randomly. The string of 1s and 0s could be read by watching how the polymer broke apart inside a mass spectrometer.

Earlier this month, Lutz showed that a library of different polymer strands could encode a 32-bit message[14]. That pales by comparison with the 1.6 gigabits that have been stored in artificial DNA molecules (see go.nature.com/2b2ve0u). But momentum is growing for polymer data storage. In April, the Intelligence Advanced Research Projects Activity (IARPA), a US agency that funds high-risk research for the intelligence community, drew representatives from the biotechnology, semiconductor and software industries to a workshop on the subject. "There's a vibrant and growing community of researchers working on this," says David Markowitz, a technical adviser at IARPA who helped to organize the workshop.

But the approach still faces enormous technical challenges: current synthetic techniques are much too slow and expensive. The key to cracking the data-storage problem — and many other problems at the polymer frontier — will be to develop better ways to predict the properties of polymers and fine-tune their production. That will require a concerted effort. "We need to establish collaborations with physicists, materials scientists, theoretical chemists," says Lutz. "We need to build a new field." ■ **SEE NEWS FEATURE PAGE 263, AND NEWS & VIEWS PAGE 276**

Mark Peplow *is a science journalist based in Cambridge, UK.*

1. Staudinger, H. *Ber. Dtsch. Chem. Ges.* **53,** 1073–1085 (1920).
2. *The New Plastics Economy: Rethinking the Future of Plastics* (Ellen MacArthur Foundation, 2016); available at go.nature.com/2bdyxep
3. Li, T., Zhang, J., Schneiderman, D. K., Francis, L. F. & Bates, S. F. *ACS Macro Lett.* **5,** 359–364 (2016).
4. Schneiderman, D. K. *et al. ACS Macro Lett.* **5,** 515–518 (2016).
5. Schneiderman, D. K. & Hillmyer, M. A. *Macromolecules* **49,** 2419–2428 (2016).
6. Lin, N. & Dufresne, A. *Eur. Polymer J.* **59,** 302–325 (2014).
7. Miller, D. J. *et al. J. Membrane Sci.* **437,** 265–275 (2013).
8. Kamcev, J. *et al. Phys. Chem. Chem. Phys.* **18,** 6021 (2016).
9. Kissel, P., Murray, D. J., Wulftange, W. J., Catalano, V. J. & King, B. T. *Nature Chem.* **6,** 774–778 (2014).
10. Kory, M. J. *et al. Nature Chem.* **6,** 779–784 (2014).
11. Barnes, J. C. *et al. Nature Chem.* **7,** 810–815 (2015).
12. Leibfarth, F. A., Johnson, J. A. & Jamison, T. F. *Proc. Natl Acad. Sci. USA* **112,** 10617–10622 (2015).
13. Roy, R. K. *et al. Nature Commun.* **6,** 7237 (2015).
14. Laure, C., Karamessini, D., Milenkovic, O., Charles, L. & Lutz, J.-F. *Angew. Chem. Int. Ed. Engl.* http://dx.doi.org/10.1002/anie.201605279 (2016).

nature International weekly journal of science

NATURE | COMMENT

Rethink how chemical hazards are tested

John C. Warner[1] & Jennifer K. Ludwig[2]

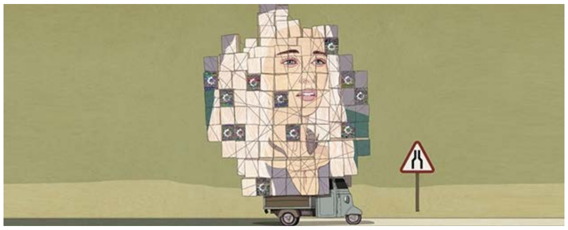
16 August 2016

John C. Warner and Jennifer K. Ludwig propose three approaches that would help inventors to produce safer chemicals and products.

Rights & Permissions

Subject terms: Chemistry • Policy

Internet winter is coming



The bandwidth bottleneck that is throttling the Internet

Researchers are scrambling to repair and expand data pipes worldwide — and to keep the information revolution from grinding to a halt.

Recent

1. **US personalized-medicine industry takes hit from Supreme Court**
   Nature | 17 August 2016
2. **CRISPR's hopeful monsters: gene-editing storms evo-devo labs**
   Nature | 17 August 2016
3. **Ethics: Taming our technologies**
   Nature | 17 August 2016

Read

1. **Replications, ridicule and a recluse: the controversy over NgAgo gene-editing intensifies**
   Nature | 08 Aug 2016

Kevin Frayer/Getty

A farmer spays pesticide on an apple tree in Hanyuan, China.

Around the world, safety regulations are being revised as new information about the health and environmental effects of chemicals becomes available. In June, US President Barack Obama signed the first bill to reform the Toxic Substances Control Act since its enactment 30 years ago. The revised act mandates greater public transparency and the timely assessment of existing chemicals by the US Environmental Protection Agency (EPA). Elsewhere, the European Union's REACH (registration, evaluation, authorization and restriction of chemicals) legislation and similar laws are also evolving.

Improved regulation is necessary to protect people and the environment from harmful substances. But it does little for inventors who face the perplexing task of creating safer chemicals and products[1]. In the current system, safety information is gathered after a chemical is invented, or in many cases, after it is incorporated into products and distributed to the public. The molecular interactions of chemicals within products are unaccounted for, meaning that ingredients lists may be misleading as sources for product safety information. Such factors make it nearly impossible for an inventor to avoid the risk of creating an unsafe chemical or product.

The evaluation and communication of chemical and product safety needs to change. Three approaches are proposed here to start a conversation between scientists, business representatives and policymakers about our future public and environmental health.

2. **E-mails show how UK physicists were dumped over Brexit**
Nature | 05 Aug 2016

3. **Beyond CRISPR: A guide to the many other ways to edit a genome**
Nature | 08 Aug 2016

View all ▸

Commented

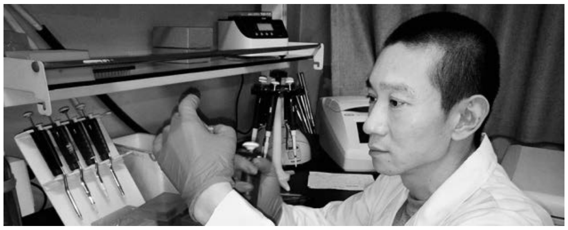
1. **Donald Trump's appeal should be a call to arms**
Nature | 03 Aug 2016 | 33 comments

2. **A safe place for nuclear energy?**
Nature | 09 Aug 2016 | 14 comments

3. **Trump vs Clinton: worlds apart on science**
Nature | 26 Jul 2016 | 13 comments

View all ▸

Replications, ridicule and a recluse



Three ways forward

Standardize chemical-safety tests. Controversy on chemical safety often arises when organizations, from corporations to research centres and government agencies, test the same compound using different methods. One technique may suggest that a compound is hazardous, another that it is benign. For example, glyphosate, a widely used herbicide, was in 2015 deemed a "probable human carcinogen" by the International Agency for Research on Cancer[2]. Many other regulatory agencies, including the European Food Safety Authority, conversely concluded that the herbicide was "unlikely to be carcinogenic". The discrepancy lies in the different studies taken as evidence, which leaves the public more confused about the safety of glyphosate than before.

Standardized tests reduce the use of replacement chemicals that are as problematic as, or worse than, the original substance. For example, some structural analogues of bisphenol A (BPA), which are used in a variety of plastic products, have similar toxicity and hormonal effects to BPA[3]. Likewise, hydrofluorocarbons and hydrochlorofluorocarbons are often used as substitutes for chlorofluorocarbons (CFCs), ozone-depleting chemicals that were used widely as refrigerants and aerosol propellants. Although not as harmful as CFCs, the substitutes still damage Earth's ozone layer[4].

Further, by knowing which tests must be carried out in advance, inventors will save time and money, making it easier to rationalize the large investment necessary to develop a material.

> **"The molecular interactions of chemicals within products are unaccounted for."**

Creating a set of nationally or internationally standardized safety tests will require input and compromise from industrial, academic and governmental organizations, such as the American Chemistry Council, the Environmental Working Group and the EPA. Everyone will endorse some tests, such as those for physical chemical properties. Others will be difficult to agree on or are yet to be established, such as those for endocrine disruptors, a type of hormone-mimicking molecule[5]. Information gaps will need to be identified, such as methodologies for testing the various phases of materials. A mechanism to periodically review and amend the list of tests should be put in place, based on existing processes for evaluating individual molecules used by the

The controversy over NgAgo gene-editing intensifies

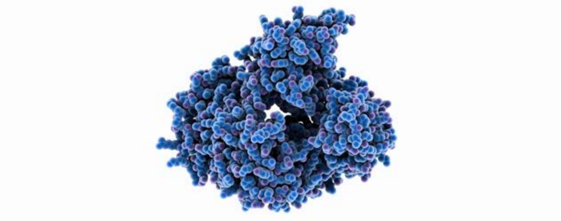
As failures to replicate results using the CRISPR alternative stack up, a quiet scientist stands by his claims.

Long in the tooth

Near-blind shark is world's longest-lived vertebrate

Greenland shark found to be at least 272 years old.

Expanded editing



Beyond CRISPR: A guide to the many other ways to edit a genome

The popular technique has limitations that have sparked searches for alternatives.

Exclusion zone

E-mails show how UK physicists were dumped over Brexit

Researchers dropped from EU grant proposal because UK inclusion would 'compromise' project.

Nature Podcast

EPA, REACH, corporations and government bodies.

Test finished products. Ingredients entering a manufacturing process do not necessarily represent the chemical composition of the final product. Some molecules disappear; others interact to form new compounds when exposed to different substances or changes in temperature and pressure. A better way to understand a product's impact on human health and the environment is to test the final product. For example, one study that screened a sample of pizza box[6] revealed many unidentifiable compounds, raising questions about the content and safety of everyday products.

A product could be graded on a scale of 1 to 10 (1 being benign and 10 being highly toxic) based on its performance in a series of standard tests in different categories. Consumers would be informed of product safety and suppliers need not reveal trade secrets. If a product's performance in one or more of the tests is unacceptable, the manufacturer can look down its supply chain, identify which material is problematic, and make modifications.

Make test results public. The quantitative results of chemical and product tests should be disclosed and presented in an unbiased way. Organizations, including government agencies, non-governmental organizations and trade associations should create policies and processes to interpret the data. For example, a product might be scored for carcinogenicity, emissions and endocrine-disrupting potential. If all products in a commercial category provide this information, a consumer can make an informed decision by comparing the numbers. Consumer or non-governmental organizations should prepare guidelines on what scores one should look for.

It is important to ensure consumers know that no product is without risk. Producers with 'unacceptable' product scores would have to explain to the public why they feel that the exposure of humans and the environment to a substance is justified. Government agencies and other groups can ban products or product categories that score poorly.

Path to progress

The first step towards better chemical safety is to create a list of desired endpoints — the information we would like to know about a product, such as liver toxicity, ozone depletion or carcinogenicity. There shouldn't be so many goals that the task of achieving them is impossible, or so few that it is meaningless.

Step two is to identify specific tests for each endpoint. Where consensus cannot be achieved, a mechanism for reaching agreement must be developed.

Third, we must develop protocols to define sample preparation and methods of analysis. The main goal is to create criteria that can be used to audit laboratories that perform the assays. Different states of matter and various product types should be anticipated.

Finally, scientists should convene regularly to evaluate the current state of the art and science, and make decisions based on new knowledge that challenges existing tests or offers improvements. For example, this year marks the twentieth anniversary of the first Green Chemistry Gordon conference; such meetings would be good forums for discussing commercial successes and remaining challenges in sustainable chemistry.

Overhauling chemical regulation is a daunting task, but we need a better way of protecting human health and the environment.

Nature **536,** 269–270 (18 August 2016) | doi:10.1038/536269a

Tweet Follow @NatureNews

References

1. Anastas, P. T. & Warner, J. C. *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
2. Guyton, K. Z. *et al*. *Lancet Oncol.* **16**, 490–491 (2015). Article PubMed
3. Rochester, J. R. & Bolden, A. L. *Environ. Health Perspect.* **123**, 643–650 (2015). PubMed
4. UNEP. *HFCs: A Critical Link in Protecting Climate and the Ozone Layer* 36 (UNEP, 2011).
5. Schug, T. T. *et al*. *Green Chem.* **15**, 181–198 (2013).

Article PubMed ChemPort

6. Bengtström, L. *et al*. *Food Addit. Contam. Part A* **33**, 1080–1093 (2016).

Article ChemPort

Related stories and links

From nature.com

- **Policy: Five cornerstones of a global bioeconomy**
  12 July 2016
- **Validate personal air-pollution sensors**
  06 July 2016
- **Industrial research: Drug companies must adopt green chemistry**
  31 May 2016
- **Chemistry: Develop ionic liquid drugs**
  08 December 2015
- **Sustainability: Don't waste seafood waste**
  10 August 2015

From elsewhere

- **European Chemicals Agency: Plans and reports**
- **2016 Global Chemical Regulations Outlook**
- **Glyphosate: EFSA updates toxicological profile**
- **TSCA Work Plan Chemicals**

Author information

Affiliations

John C. Warner is president and chief technology officer, at the Warner Babcock Institute for Green Chemistry, Wilmington, Massachusetts, USA.

Jennifer K. Ludwig is a scientific technical writer, at the Warner Babcock Institute for Green Chemistry, Wilmington, Massachusetts, USA.

Corresponding authors

Correspondence to: John C. Warner or Jennifer K. Ludwig

Author details

John C. Warner

Contact John C. Warner

Jennifer K. Ludwig

a chemical is invented, or in many cases, after it is incorporated into products and distributed to the public. The molecular interactions of chemicals within products are unaccounted for, meaning that ingredients lists may be misleading as sources for product safety information. Such factors make it nearly impossible for an inventor to avoid the risk of creating an unsafe chemical or product.

The evaluation and communication of chemical and product safety needs to change. Three approaches are proposed here to start a conversation between scientists, business representatives and policymakers about our future public and environmental health.

THREE WAYS FORWARD

Standardize chemical-safety tests. Controversy on chemical safety often arises when organizations, from corporations to research centres and government agencies, test the same compound using different methods. One technique may suggest that a compound is hazardous, another that it is benign. For example, glyphosate, a widely used herbicide, was in 2015 deemed a "probable human carcinogen" by the International Agency for Research on Cancer[2]. Many other regulatory agencies, including the European Food Safety Authority, conversely concluded that the herbicide was "unlikely to be carcinogenic". The discrepancy lies in the different studies taken as evidence, which leaves the public more confused about the safety of glyphosate than before.

Standardized tests reduce the use of replacement chemicals that are as problematic as, or worse than, the original substance. For example, some structural analogues of bisphenol A (BPA), which are used in a variety of plastic products, have similar toxicity and hormonal effects to BPA[3]. Likewise, hydrofluorocarbons and hydrochlorofluorocarbons are often used as substitutes for chlorofluorocarbons (CFCs), ozone-depleting chemicals that were used widely as refrigerants and aerosol propellants. Although not as harmful as CFCs, the substitutes still damage Earth's ozone layer[4].

Further, by knowing which tests must be carried out in advance, inventors will save time and money, making it easier to rationalize the large investment necessary to develop a material.

Creating a set of nationally or internationally standardized safety tests will require input and compromise from industrial, academic and governmental organizations, such as the American Chemistry Council, the Environmental Working Group and the EPA. Everyone will endorse some tests, such as those for physical chemical properties. Others will be difficult to agree on or are yet to be established, such as those for endocrine disruptors, a type of hormone-mimicking molecule[5]. Information gaps will need to be identified, such as methodologies for testing the various phases of materials. A mechanism to periodically review and amend the list of tests should be put in place, based on existing processes for evaluating individual molecules used by the EPA, REACH, corporations and government bodies.

Test finished products. Ingredients entering a manufacturing process do not necessarily represent the chemical composition of the final product. Some molecules disappear; others interact to form new compounds when exposed to different substances or changes in temperature and pressure. A better way to understand a product's impact on human health and the environment is to test the final product. For example, one study that screened a sample of pizza box[6] revealed many unidentifiable compounds, raising questions about the content and safety of everyday products.

> **"The molecular interactions of chemicals within products are unaccounted for."**

A product could be graded on a scale of 1 to 10 (1 being benign and 10 being highly toxic) based on its performance in a series of standard tests in different categories. Consumers would be informed of product safety and suppliers need not reveal trade secrets. If a product's performance in one or more of the tests is unacceptable, the manufacturer can look down its supply chain, identify which material is problematic, and make modifications.

Make test results public. The quantitative results of chemical and product tests should be disclosed and presented in an unbiased way. Organizations, including government agencies, non-governmental organizations and trade associations should create policies and processes to interpret the data. For example, a product might be scored for carcinogenicity, emissions and endocrine-disrupting potential. If all products in a commercial category provide this information, a consumer can make an informed decision by comparing the numbers. Consumer or non-governmental organizations should prepare guidelines on what scores one should look for.

It is important to ensure consumers know that no product is without risk. Producers with 'unacceptable' product scores would have to explain to the public why they feel that the exposure of humans and the environment to a substance is justified. Government agencies and other groups can ban products or product categories that score poorly.

PATH TO PROGRESS

The first step towards better chemical safety is to create a list of desired endpoints — the information we would like to know about a product, such as liver toxicity, ozone depletion or carcinogenicity. There shouldn't be so many goals that the task of achieving them is impossible, or so few that it is meaningless.

Step two is to identify specific tests for each endpoint. Where consensus cannot be achieved, a mechanism for reaching agreement must be developed.

Third, we must develop protocols to define sample preparation and methods of analysis. The main goal is to create criteria that can be used to audit laboratories that perform the assays. Different states of matter and various product types should be anticipated.

Finally, scientists should convene regularly to evaluate the current state of the art and science, and make decisions based on new knowledge that challenges existing tests or offers improvements. For example, this year marks the twentieth anniversary of the first Green Chemistry Gordon conference; such meetings would be good forums for discussing commercial successes and remaining challenges in sustainable chemistry.

Overhauling chemical regulation is a daunting task, but we need a better way of protecting human health and the environment. ■

John C. Warner *is president and chief technology officer, and* **Jennifer K. Ludwig** *is a scientific technical writer, at the Warner Babcock Institute for Green Chemistry, Wilmington, Massachusetts, USA.*
e-mails: john.warner@warnerbabcock.com; jennifer.ludwig@warnerbabcock.com

1. Anastas, P. T. & Warner, J. C. *Green Chemistry: Theory and Practice* (Oxford Univ. Press, 1998).
2. Guyton, K. Z. *et al. Lancet Oncol.* **16,** 490–491 (2015).
3. Rochester, J. R. & Bolden, A. L. *Environ. Health Perspect.* **123,** 643–650 (2015).
4. UNEP. *HFCs: A Critical Link in Protecting Climate and the Ozone Layer* 36 (UNEP, 2011).
5. Schug, T. T. *et al. Green Chem.* **15,** 181–198 (2013).
6. Bengtström, L. *et al. Food Addit. Contam. Part A* **33,** 1080–1093 (2016).

CORRECTION

The Comment article 'Stop the privatization of health data' (J. T. Wilbanks & E. J. Topol *Nature* **535,** 345–348; 2016) wrongly stated that the Enlite device sends insulin into the blood when it detects a drop in glucose; in fact, it stops a pump releasing insulin. And 23andMe's latest fundraising round was US$115 million, not $150 million.

UNIVERSAL IMA/REX/SHUTTERSTOCK

Humanoid robot iCub is used for research into cognition and artificial intelligence.

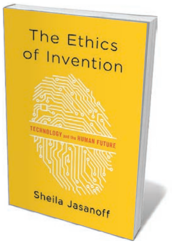
ETHICS

Taming our technologies

Steven Aftergood weighs up a study that gauges the gap between oversight and the onward rush of innovation.

Technological innovation in fields from genetic engineering to cyberwarfare is accelerating at a breakneck pace, but ethical deliberation over its implications has lagged behind. Thus argues Sheila Jasanoff — who works at the nexus of science, law and policy — in *The Ethics of Invention*, her fresh investigation. Not only are our deliberative institutions inadequate to the task of oversight, she contends, but we fail to recognize the full ethical dimensions of technology policy. She prescribes a fundamental reboot.

Ethics in innovation has been given short shrift, Jasanoff says, owing in part to technological determinism, a semi-conscious belief that innovation is intrinsically good and that the frontiers of technology should be pushed as far as possible. This view has been bolstered by the fact that many technological advances have yielded financial profit in the short term, even if, like the ozone-depleting chlorofluorocarbons once used as refrigerants, they have proved problematic or ruinous in the longer term.



The Ethics of Invention: Technology and the Human Future
SHEILA JASANOFF
W. W. Norton: 2016

There are contingent issues. Numerous government and professional ethical advisory bodies already exist, looking at research with human subjects and specific fields of innovation. But they tend to have a technocratic orientation that focuses on cost–benefit analysis narrowly construed, with an emphasis on those factors that can be quantified or assigned market value. Intangibles, such as worker morale or the health of communities, are often neglected. Meanwhile, technological disasters such as the 2010 Deepwater Horizon oil spill in the Gulf of Mexico lay bare flaws in the conception, design or implementation of technology, at least after the fact. But because such failures are by definition unintended, they are typically exempted from deep ethical concern by planners and regulators.

What we too often fail to grapple with, writes Jasanoff, is that technology is value-laden from start to finish. From the innovator's intuition of a desired end to the development of the practical means of achieving that end — as well as its application, distribution, ownership and ultimate impact on society and the world at large — choices about technology are inextricably intertwined with value judgements at every stage.

Jasanoff argues for an entirely new body of ethical discourse, going beyond technical risk assessment to give due weight to economic, cultural, social and religious perspectives. *The Ethics of Invention* is an eloquent meditation on these problems. Jasanoff thoughtfully discusses the limits of conventional risk analysis, with its biases in favour of innovation and quantification, and looks at the challenges posed by specific developments in biotechnology, genetic engineering and information technology for surveillance. The book helps to pinpoint recurring patterns in contemporary technological debates and to frame what is at stake in their outcomes.

But the solution to the "deep democratic deficit" in current technology policy is hard to articulate and harder still to implement, and Jasanoff does not provide a clear road map. An attempt to take all relevant ethical perspectives into account may be a prescription for stalemate because, as she notes, "many basic issues of right and wrong remain deeply contested". These include questions such as when life begins and ends, what constitutes human dignity and how the scope of human responsibility to the global environment and to future generations can be defined. Sometimes, intensive ethical deliberation leads not to consensus, but to its opposite. After years of study and public debate, for example, the 1984 industrial disaster in Bhopal, India — when toxic gases leaked from a Union Carbide pesticide plant, killing thousands — now "stubbornly resists any coherent narrative of causes and effects" owing to complex political, legal and jurisdictional conflicts. At other times, deliberation can yield multiple, disparate concurrences: different policies on human embryonic-stem-cell research have been adopted by the United States, the United Kingdom and Germany, for instance.

Yet even if an ideal ethical discourse is out of reach, Jasanoff persuasively argues,

▶

▶ we can do better than we have done. The impacts of many technologies in fields such as energy production, robotics or knowledge management extend far beyond their operators or beneficiaries, so it is necessary to find a way to solicit and consider the views of affected populations. Scientists increasingly recognize this (J. Kuzma *Nature* **531,** 165–167; 2016). A reliance on technologists alone, the author opines, would be an error because "experts' imaginations are often circumscribed by the very nature of their expertise". The US congressional Office of Technology Assessment, closed in 1995, used to independently evaluate a wide range of technology problems. Now viewed with nostalgia by critics who lament the scientific illiteracy of much of contemporary politics, it receives only qualified admiration from Jasanoff. She concludes that, in several cases, the organization "failed to carve out the space of neutral expertise that its designers had hoped for" and instead "became one more loud, discordant note in the ongoing cacophonous debate".

Some of the most intriguing portions of the book deal with the personally transformative effects of technology. "Our inventions change the world, and the reinvented world changes us," as Jasanoff puts it. Technology determines our sense of the possible and can enhance or diminish our natural abilities, even altering brain size and function (R. McKinlay *Nature* **531,** 573–575; 2016). Our technological choices are both reflections of who we are and stepping stones towards who we will become: emerging technologies may yet redefine what it means to be human. Depending on what we value most — power, knowledge, sustainability, conviviality or convenience — some technologies will serve us well and others must be excluded. Ethics is central to the process of choosing between them.

"Experts' imaginations are often circumscribed by the very nature of their expertise."

Expanding the scope of ethical deliberation over new technology may seem like a daunting prospect bound to impede innovation. It will undoubtedly raise questions more quickly than they can be answered. But experience suggests that many such questions will be worth asking. ■

Steven Aftergood *is a senior research analyst at the Federation of American Scientists in Washington DC.*
e-mail: saftergood@fas.org

Q&A Brenda Keneghan

The polymer conservator

For many, plastic is a dirty word — a pollutant that can't degrade soon enough. But for polymer scientist Brenda Keneghan, it's a precious material that looms large in design history. A conservator at the Victoria and Albert Museum (V&A) in London, Keneghan spends her days saving plastic items from furniture to toys from the ravages of time. Here she talks about the war against the warping, yellowing, crumbling and stickiness that plague polymers.

When did people first recognize that plastic degradation was a problem?

For a long time, no one noticed that plastics were degrading, because they were used for throwaway objects. But by the 1960s, a cellulose acetate sculpture by Russian constructivist artist Naum Gabo, at the Philadelphia Museum of Art in Pennsylvania, was disintegrating. Its value made people take notice. When I joined the V&A in 1994, some people still saw plastic as an imitative material rather than a material in its own right. Now, they respect it as a medium that can take any shape, that is used to create objects and artworks that you couldn't make any other way.

What is in the V&A's plastic collection?

We have shoes, accessories and bags from the 1920s, when manufacturers used cellulose acetate or nitrate to imitate natural materials such as amber, ivory and tortoiseshell, for example in intricate hair combs. In the 1930s and 1940s, couture designers including Elsa Schiaparelli played with the new materials, which by then also included semi-synthetics such as casein (made from milk protein and formaldehyde), in decorations, buttons and fabrics. From the 1960s, pop-art furniture made of polyurethane (PU) and other plastics emerged, including a sleek single-mould chair by Danish designer Verner Panton and inflatable furniture. There were also polyvinyl chloride (PVC) raincoats and boots. The V&A has beautiful radios and cigarette cases made of Bakelite, but that is pretty stable. In our outpost at the Museum of Childhood in London, we have a huge range of plastic toys, including PVC dolls and PU foam figures.

How do plastics degrade?

Only five types degrade catastrophically in reaction to humidity, light and air — it's a problem of thermodynamics. It can take from a few years to a few decades. Cellulose acetate and nitrate react with moisture in the air and crumble, producing acid vapours that can corrode anything that shares their display cases. For PU, the problem is oxidation: once additives such as phenolic antioxidants are used up, the plastic crumbles. (The Museum of Childhood's PU foam puppets of television character Larry the Lamb have succumbed completely.) PVC degrades because of the plasticizer molecules that make it flexible. These sit in the mixture and creep up to the surface, making it sticky. Both plasticizers and the base polymer can undergo a reaction that makes the surface dark. Finally, natural rubber, which filled the cushions of pre-1950s upholstery and formed shoe soles, will oxidize and become brittle over time. For all of these, we can stave off the process through specialized storage and display conditions, but we can't prevent it entirely.

What happens when the museum acquires a new plastic object?

We have to weigh up the care and cost demanded by the plastic's type, age and condition. Many older objects are not labelled, so you need to discover what plastics they are made of. We use an infrared microscope to find the material's fingerprint — how it absorbs different wavelengths of light. Smell and appearance are important, but not conclusive. Our intern Carien van Aubel is looking for features in dozens of test objects to spot similarities — perhaps production techniques or usage — that would help museums without our testing facilities to identify polymers in new acquisitions. This also applies when we put together touring exhibitions: we need to know whether the items we borrow can withstand the rigours of transport and display. In 2013, for example, we checked costumes for the V&A's

WWW.LUNETIERVINTAGE.COM

Plastic butterfly sunglasses from the 1960s.

exhibition David Bowie Is, including PVC boots and a skirt-like structure stiffened with polyurethane foam.

Are there any V&A pieces that are ticking time bombs?
I wouldn't say we have anything that dramatic. Much of our 3D-printed furniture is white nylon, which can yellow. We have a little cellulose acetate box by early-twentieth-century French glassware designer René Lalique that is definitely warping. The problem is worse for modern-art museums, because they have many contemporary artworks that incorporate plastics. From the 1980s, the number of plastics available exploded, and people largely stopped using the least stable ones. But plastics do not last forever, because — unlike wood, metal or stone — they can be damaged by adhesives used for repair. So even something made of Perspex might last only a century, although we don't really know because the material hasn't been around for that long.

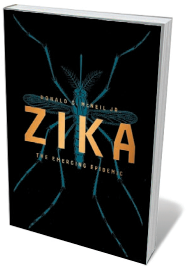
What brought you to the V&A?
I have always been interested in collecting old objects, but mainly plastic items that you would pick up in junk shops, such as Bakelite cigarette boxes. My first degree is in chemistry and my PhD in materials science, focusing on polymers; so when this job came up, it was nice to marry the science with art and design.

Do artists and manufacturers factor plastics degradation into their work?
No. A lot of modern art is made of plastic, and often artists, aiming for a particular effect, ignore the manufacturer's instructions on how to mix it. So the plastic might degrade even more quickly. We also now have bioplastics, which are made to degrade. With these, you'll definitely be fighting against the tide. ■ SEE NEWS FEATURE P.266

INTERVIEW BY ELIZABETH GIBNEY

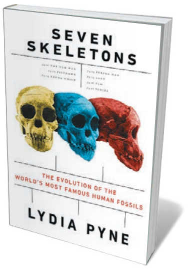
This interview has been edited for length and clarity.

Books in brief

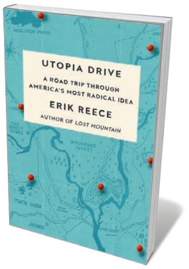


Zika: The Emerging Epidemic
Donald G. McNeil Jr W. W. Norton *(2016)*
Zika is strangely anomalous. In 99% of cases, the symptoms of this mosquito-borne or sexually transmitted flavivirus are mild, but it can wreak havoc in fetuses, crossing the placenta to trigger brain defects such as microcephaly. In this agile account, science reporter Donald McNeil covers Zika's discovery in 1940s Uganda, early cases, the Brazilian outbreaks of 2015 and the implications of the virus's spread. McNeil's mapping of official responses to the epidemic, from early statements that Zika was benign to recognition of its virulence and the race towards a vaccine, underlines the burning need for viral vigilance.

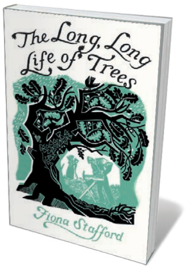
Sleep in Early Modern England
Sasha Handley Yale University Press *(2016)*
Sleep became a hotbed of speculation and science in early-modern England, reveals historian Sasha Handley in this absorbing study. A complex cultural phenomenon viewed as a fluid midpoint "on the path of transformation between life and death", it also became a proving ground for advances in physiology. So Thomas Willis — who mapped blood flow between brain and body — identified the nervous system as central to sleep regulation. Such findings fed into technologies such as nightcaps to 'protect' the brain, as well as elaborate dream theories linking blood stagnation and nightmares.



Seven Skeletons
Lydia Pyne Viking *(2016)*
Why do certain scientific discoveries gain celebrity? Lydia Pyne teases apart the histories of seven hominin fossils to find out. She reveals how a virtuosic 1911 description by palaeontologist Marcellin Boule helped to make the Old Man of La Chapelle, a nearly complete Neanderthal fossil, a species archetype and cultural icon for decades. She shows how 3.5-million-year-old australopith Lucy became a research benchmark and a world-touring superstar. And she reminds us how these nodes in our storied past grip our collective imagination even as they add immeasurably to evolutionary science.



Utopia Drive: A Road Trip Through America's Most Radical Idea
Erik Reece Farrar, Straus and Giroux *(2016)*
From 1820 to the 1850s, the US east coast heaved with social experiments, as citizens disaffected by socio-economic turmoil "plotted paradise" in nearly 200 utopian communities, only to disperse with the Civil War. Erik Reece's meander through a number of sites takes in the philosophical roots of the Shakers' sublimely ingenious designs; the lab-like confines of Walden, where Henry David Thoreau conducted his microeconomics trial; and the potential, in our globalized, hyper-consuming era, for communal economies, land trusts and other utopian solutions to re-root.



The Long, Long Life of Trees
Fiona Stafford Yale University Press *(2016)*
In this paean to the "arboreal impulse", Fiona Stafford gets under the bark of the terrestrial giants whose natural history is interlaced with our own. Interspersed with crisp black-and-white illustrations, Stafford's low-down on species from ash to yew mesh dendrology with cultural biography and pack in the facts — from 40,000-year-old evidence of olives on the Greek island of Santorini to how willows are "naturally flirtatious, cross-pollinating compulsively" and why pine forests create their own cloud cover. Barbara Kiser

Correspondence

Royal Society helps guide Brexit science

While the United Kingdom's relationship with the European Union is in flux, I wish to emphasize that the Royal Society's president, Venki Ramakrishnan, and its foreign secretary, Martyn Poliakoff, are more strongly engaged with Europe than ever. They are determined to continue as an effective voice for international, collaborative science (see *Nature* **535,** 467; 2016).

Prime Minister Theresa May said she is committed to "ensuring a positive outcome for UK science as we exit the European Union". The Royal Society will work with the government to turn these words into actions and to see that the UK science community is heard in Brexit negotiations.

We have already laid the groundwork to help the country maintain its world-leading position in research and to continue attracting the best international researchers. We are also using evidence that was gathered by the Royal Society before the 23 June EU referendum to identify the best possible model of international collaboration to pursue within and beyond the EU.

Last month, our president called on the government to address post-Brexit uncertainty by underwriting the research of all UK-based scientists funded by the EU. Under #ScienceIsGlobal, we have brought more than 30 organizations from 24 European nations together to seek assurance from their respective governments that our scientists can continue to collaborate for the benefit of society.

Julie Maxton *Royal Society, London.*
public.affairs@royalsociety.org

Being more open about PhD papers

Submitting a PhD thesis as a compilation of research papers can help scientists' early careers (see *Nature* **535,** 26–28; 2016), but acknowledgements and declarations should not be overlooked along the way.

In the Netherlands, a PhD student's research articles — often co-authored by the supervisor — are sandwiched between introductory and concluding chapters. The thesis is published before the viva voce exam with an ISBN identifier and is later posted online. Advantages over the traditional monograph thesis include: it is quick and easy to write; feedback from the papers' reviewers can be instructive; and students attain a presence in the international science community before graduation.

I suggest that, out of courtesy, people involved in the publishing process should be informed that the papers will be assessed as part of a higher degree. They include journal reviewers and editors, as well as language professionals like me who are asked to correct the English of the manuscripts. The thesis itself could also contain a prominent statement of all assistance received, along with a declaration of the candidate's input (see B. Gustavii *How to Prepare a Scientific Doctoral Dissertation Based on Research Articles*; Cambridge Univ. Press, 2012).

Joy Burrough-Boenisch *Renkum, the Netherlands.*
unclogged.english@gmail.com

World's last *in vitro* fertilization ban falls

After a lengthy struggle, *in vitro* fertilization (IVF) procedures began last month in Costa Rica. This effectively ends the last full IVF ban in the world. (In countries under Islamic law, for example, IVF is permitted, albeit only within marriage.)

IVF was banned in 2000 in Costa Rica, one of the few remaining countries where Catholicism is the state religion. The ban followed pressure from religious extremists to limit women's rights and to claim full personhood for fertilized eggs (zygotes). Members of the legal, medical and scientific communities countered vigorously, for example by pointing out that a zygote remains incomplete without the developmental signals that result from implantation in the womb. Humans, after all, are not oviparous — we do not lay eggs.

The Inter-American Court of Human Rights invalidated the ban in 2012 in support of those seeking IVF treatment. However, extremist lawmakers continued to obstruct IVF until February, when the court nullified their attempts to block an executive order regulating it.

Nevertheless, the battle for women's rights in Costa Rica is far from over. For example, the country has yet to implement existing legislation that allows abortion when maternal health is compromised.

Felipe Mora-Bermúdez
Max Planck Institute of Molecular Cell Biology and Genetics, Dresden, Germany.
mora@mpi-cbg.de

Stop vultures from striking aircraft

An ecological solution is needed to prevent collisions of Eurasian griffon vultures (*Gyps fulvus*) with aircraft in Spain, home to some 95% of Europe's population of these large raptors. There were 26 such collisions recorded in 2006–15 around Madrid Barajas airport, which handles about 47 million passengers each year, and 3 light-aircraft strikes in the first 6 months of 2016.

This surge could be explained by the birds' relocation away from their usual feeding areas, following changes in European health regulations in 2002 (see A. Margalida *et al. Nature* **480,** 457; 2011). Those rules forced farmers to collect and destroy livestock carcasses, an important food source for vultures.

The birds have since scavenged in areas such as landfill sites. When these are located inside air-traffic corridors (where aircraft fly below 1,200 metres), the risk of collision increases. The negative effects of the 2002 health regulations on vulture populations and demography are being tackled (see go.nature.com/2ap6zsd), but not fast enough to avert aerial collisions.

To manage the situation more effectively, we need a better understanding of the movement ecology of vultures around sensitive areas, the spatial distribution of their food resources and a warning system that detects vultures entering air-traffic corridors.

Antoni Margalida *University of Lleida, Spain.*
amargalida@ca.udl.cat

Music calculations out of tune

It is perhaps not surprising that the Tsimane' villagers in Bolivia cannot tell the difference between minor and major keys, or dissonant and non-dissonant sounds (*Nature* **535,** 199–200; 2016). Alternative musical scales and intervals have been known to musicians for centuries — for example, they often account for the varying timbres of different folk music (see also go.nature.com/2avI1fn).

Your assertion that "consonant intervals … are integral ratios of harmonics — 2:1, 4:3 and 3:2" is out of tune. Except for octaves (2:1), this has not been true in Western music since at least Bach's time. The title of his famous collection of fugues, 'The Well-Tempered Clavier', does not allude to his clavier's behaviour. All modern pianos, and therefore modern Western music, are in fact equal-tempered, such that the increase in frequency per semitone is $2^{1/12}$ — that is, about 1.059 (see go.nature.com/2aazsvs). For each key from C major to B minor, the perfect fifth is not a ratio of 1:1.5, but 1:1.498.

Adrian Goldman *University of Leeds, UK.*
a.goldman@leeds.ac.uk

In retrospect

Sixty years of living polymers

In the 1950s, the discovery of a class of 'living' polymerization reaction revolutionized the field of polymer science by providing a way of controlling the molecular-weight distribution of polymers. The effects reverberate to this day.

GARY PATTERSON

One of the triumphs of modern polymer science is the exquisite control that synthetic chemists have achieved in the design and execution of polymerization reactions[1]. A key concept on which this control is based was discussed 60 years ago by Michael Szwarc[2] in a classic paper in *Nature*. He reported 'living' polymerization reactions, in which each addition of a monomer to a growing chain is irreversible and, when the pool of monomers is exhausted, the ends of the polymer chain remain active so that further chemistry can take place. Szwarc's findings have been applied to a wide range of polymerizations, and are responsible not only for major industrial applications, but also for advancing the theory of polymer science[3].

In the early 1950s, typical laboratory polymerizations produced a mixture of molecules of different chain lengths because the reactions were reversible — monomers could detach from polymer chains, rather than irreversibly adding to them, and random termination reactions could occur, preventing further chain growth and causing even broader chain-length distributions. Theoretical considerations suggested that many of the properties of polymers depend on both the average molecular chain length and the chain-length distribution. Polymer scientists therefore required samples that were both well characterized and of controlled length to test fundamental theories. Early efforts to carry out such evaluations required extremely tedious fractionations of polymer samples to obtain appropriate test materials. The idea that irreversible polymerizations would produce polymers that have narrow molecular-weight distributions had been proposed by the chemist Paul Flory[4] in 1940, but little progress towards such reactions was made until Szwarc's paper appeared.

Szwarc received his degree in chemical engineering from the Warsaw University of Technology in 1932, but wisely chose to emigrate to Israel in 1935, before the start of the Second World War. He received his PhD in organic chemistry in 1942 from the Hebrew University of Jerusalem. In 1945, he joined the research group of Michael Polanyi — a polymath who made great contributions to physical chemistry — at the University of Manchester, UK, earning another PhD in physical chemistry in 1947, and a DSc in 1949. He joined the faculty as a senior lecturer, but then moved to the United States in 1952 to become professor of physical chemistry and polymers at the New York State College of Forestry in Syracuse. The University of Manchester was the pre-eminent place for research in polymers in Britain during the period Szwarc was there, and he was determined to continue this research at Syracuse.

Good things happen when a truly prepared mind is exposed to an otherwise disappointing result, and so it was for Szwarc. He heard reports of an 'unwanted' polymerization reaction that occurs between the radical anion of naphthalene and the monomer styrene (a radical anion is a compound that bears both a negative charge and an unpaired electron; in this case, the electron serves as an initiator for the polymerization reaction). Further studies by Szwarc found that the initial product of this reaction is another chemically active radical anion that reacts irreversibly with more styrene to produce an intriguing polymer. This reactive polymer was indefinitely stable when stored in a dry, oxygen-free solvent, but the active chain ends could be terminated — chemically inactivated — at will by adding a little moisture. This is the kind of polymer envisaged by Flory in 1940.

The realization of a chemical route to a living polymer produced a flurry of research[5], and many different polymers with narrow molecular-weight distributions were produced. Polymer physicists (such as myself) were thrilled, because it allowed materials to be prepared that could test our theories. But Szwarc realized that synthetic organic chemists would be even more pleased, because a different monomer could be added to the

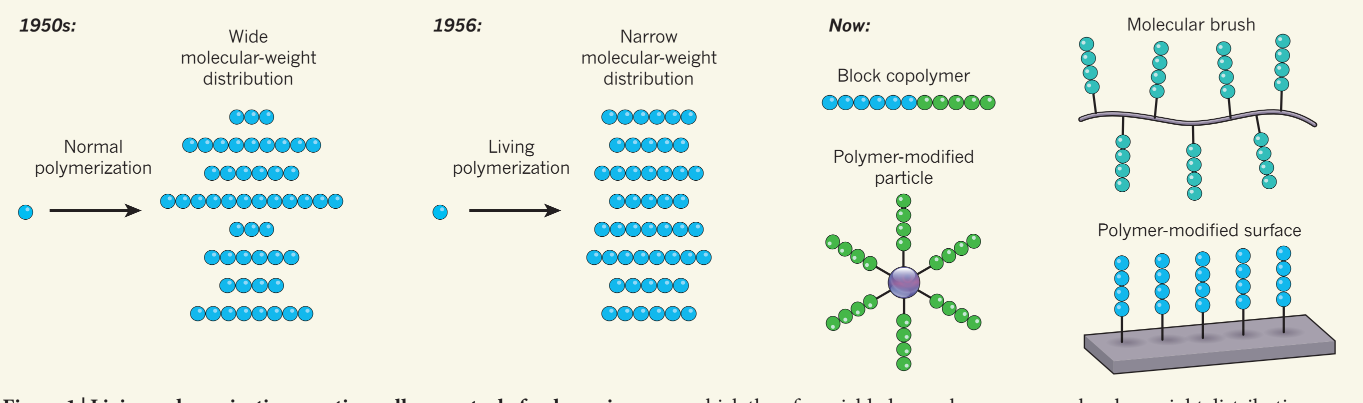


Figure 1 | Living polymerization reactions allow control of polymeric structures. In the early 1950s, most polymerization reactions produced a mixture of molecules of different chain lengths — a wide molecular-weight distribution. In 1956, Szwarc[2] reported a 'living' polymerization reaction that allowed much greater control of the products, and which therefore yielded a much narrower molecular-weight distribution. Living polymerizations have since been used to make a wide array of polymer structures, including block copolymers (which contain more than one type of monomer), molecular brushes and polymer-modified particles and surfaces.

living polymer to produce a block copolymer: molecules that contain long, uniform runs of different monomers.

Block copolymers have become major commercial successes — for example, the whole field of thermoplastic elastomers is based on this technology. Thermoplastic elastomers are rubbery solids that, unlike conventional rubbers, can be reused by heating them to temperatures above their glass transition temperature, remoulding them and then rapidly cooling them (the glass transition temperature is the range of temperatures in which amorphous materials pass from a liquid state to a hard, glassy substance). Apart from block copolymers, a dizzying number of other polymeric molecular structures engineered by living polymerization are also now available (Fig. 1). Szwarc received international recognition for the synthetic aspect of his work when he was awarded the Kyoto Prize for advanced technology in 1991.

A development that was greatly aided by the routine availability of polymers with a narrow molecular-weight distribution was the scaling theory that allows many polymer properties to be expressed in terms of molecular weight. For example, in 1950, Flory and Thomas Fox determined an equation[6] that accurately expressed the glass transition temperature as a function of molecular weight. The improved polystyrene samples available after 1956 confirmed this prediction[7].

A crucial property of pure liquid polymers is their viscosity. Flory and Fox discovered[8] that, for high-molecular-weight polymers, the viscosity increases in proportion to the molecular weight raised to the power of 3.4, and they proposed a theory to explain this finding. This means that, even well above the glass transition temperature, such polymers can have a high viscosity and behave like a soft solid. That may seem an obscure finding, but it has practical applications — such as the polymeric 'solvent' used in advanced batteries that do not leak. Again, Szwarc's discovery allowed Flory and Fox's theory to be validated.

One of the theoretically most challenging issues in scaling theory was the molecular-weight dependence of the osmotic pressure of polymer solutions. This is of interest because many industrial polymers are used in solution, and because biologists require an understanding of naturally occurring polymer solutions. The physicist Pierre-Gilles de Gennes correctly intuited[9] that, because linear polymer chains in solution are 'swollen' by the solvent, the osmotic pressure will have a different molecular-weight dependence from that predicted by classical theory. Measurements[10] of osmotic pressure for solutions encompassing wide ranges of concentration and molecular weight confirmed de Gennes' predictions. Both Flory and de Gennes received a Nobel prize for their work in polymer science and condensed-matter science, respectively.

Many theoretically challenging issues remain to be solved in polymer science, and the synergistic relationship between theory and the availability of well-defined polymer samples will greatly aid this effort. For instance, rubbery materials are widespread in industry and in biology, yet the theory of rubber elasticity is yet to be fully validated. The chemistry of living polymers also remains a highly active area[11], with hundreds of investigators worldwide. Many synthetic routes to living polymers have been developed, and a wide range of monomers can now be used in this approach. The concept of living polymers has truly revolutionized the practice of polymer science. ■

Gary Patterson *is in the Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, USA.*
e-mail: gp9a@andrew.cmu.edu

1. Matyjaszewski, K. & Müller, A. H. E. *Prog. Polym. Sci.* **31,** 1039–1040 (2006).
2. Szwarc, M. *Nature* **178,** 1168–1169 (1956).
3. Franta, E. *et al. J. Polym. Sci. A* **45,** 2576–2579 (2007).
4. Flory, P. J. *J. Am. Chem. Soc.* **62,** 1561–1565 (1940).
5. Szwarc, M. *Living Polymers and Mechanisms of Anionic Polymerization*; *Adv. Polym. Sci.* **49** (Springer, 1983).
6. Fox, T. G. & Flory, P. J. *J. Appl. Phys.* **21,** 581–591 (1950).
7. Ferry, J. D. *Viscoelastic Properties of Polymers* 3rd edn (Wiley, 1980).
8. Fox, T. G. Jr & Flory, P. J. *J. Phys. Chem.* **55,** 221–234 (1951).
9. de Gennes, P.-G. *Scaling Concepts in Polymer Physics* (Cornell Univ. Press, 1979).
10. Patterson, G. *Physical Chemistry of Macromolecules* Ch. 5 (CRC Press, 2007) .
11. Matyjaszewski, K. *Macromolecules* **45,** 4015–4039 (2012).

HUMAN GENOMICS

A deep dive into genetic variation

The exome is the portion of the genome that encodes proteins. Aggregation of 60,706 human exome sequences from 14 studies provides in-depth insight into genetic variation in humans. SEE ARTICLE P.285

JAY SHENDURE

Just seven years ago, my colleagues and I reported the protein-coding DNA sequences, called exomes, of 12 individuals[1] — among the first to be produced with a new generation of sequencing technologies[2]. Exome sequencing is much less expensive than whole-genome sequencing and, for cancers and Mendelian disorders (the latter caused by mutations in single genes), there is much more disease-associated genetic variation in the exome than in the rest of the genome. On page 285, the Exome Aggregation Consortium (ExAC) and collaborators[3] report the exome sequences of 60,706 individuals, collected from diverse studies: a venture 5,000 times larger than our initial study.

The current work highlights the pace at which human genetics is being scaled up. The project is almost ten times bigger than the Exome Sequencing Project (ESP) reported in 2013 (ref. 4), which was an important forerunner of ExAC. Indeed, this may be the deepest dive into the well of human genetic variation so far.

The study and accompanying database are noteworthy on several counts. First, for the sheer number of individuals sequenced and the depth of coverage — that is, the number of times each nucleotide in each individual's exome was sequenced. In the recently completed 1000 Genomes Project, 2,504 genomes were shallowly sequenced[5], a cost-saving strategy that favours the discovery of common over rare genetic variation. By contrast, each exome in ExAC has been sequenced deeply. Consequently, even genetic variants observed in just one individual can be confidently considered to be real (Fig. 1).

More than half of the approximately 7.5 million variants found by ExAC are seen only once. But collectively, they occur at a remarkably high density — at one out of every eight sites in the exome. For each gene, the authors contrasted the expected and observed numbers of variants that cause the production of truncated proteins, to search for regions containing lower-than-predicted levels of protein-truncating variants. This allowed them to identify several thousand genes that are highly sensitive to such variants — that is, unable to function normally after loss of one copy of the gene, even if the other copy is intact. Most of these genes have not yet been associated with disease, but mutation probably leads to embryonic death or strongly affects fitness in some other way. These genes are also intolerant of variants in regulatory DNA sequences that markedly alter levels of RNA synthesis from the gene[6], and are more likely than other genes to be implicated in genome-wide association studies of common disease.

The second noteworthy achievement of the research is that it provides a glimpse

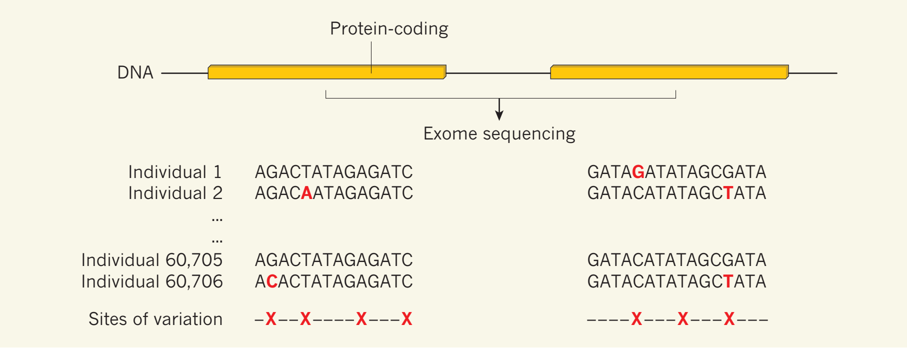

Protein-coding
DNA
Exome sequencing
Individual 1 AGACTATAGAGATC GATAGATATAGCGATA
Individual 2 AGACAATAGAGATC GATACATATAGCTATA
...
...
Individual 60,705 AGACTATAGAGATC GATACATATAGCGATA
Individual 60,706 ACACTATAGAGATC GATACATATAGCTATA
Sites of variation –X––X––––X–––X ––––X–––X–––X–––

Figure 1 | Exome aggregation. The Exome Aggregation Consortium (ExAC)[3] reanalysed the raw DNA-sequencing data from the protein-coding part of the genome, known as the exome, of 60,706 individuals, aggregated from 14 distinct studies. Genetic variants (red) are compared to produce a database of all sites of variation between the individuals.

of the bottom of the well of genetic variation in humans. In human genetics, it is generally assumed that when the same variant is found in more than one individual, it arose once in an ancestor shared by those individuals, rather than through independent mutations of the same site. However, at a particular class of site, called CpG dinucleotides, the researchers make a convincing case that variants observed in multiple individuals often reflect mutational recurrence.

In support of their assertion, the researchers find that discovery rates for new CpG dinucleotide mutations decrease in samples larger than 20,000 individuals. This provides further evidence that the size of the ExAC cohort is sufficiently large that we are beginning to saturate this class of human genetic variation, at least within the exome. It is worth noting, however, that CpG dinucleotides have a highly elevated mutation rate in human genomes, making the number of samples needed to observe such saturation much lower than for other kinds of variants. Nonetheless, this exciting finding presages what lies ahead, as larger aggregate analyses of exomes and genomes are performed.

Third, ExAC promotes the discovery of genes involved in rare diseases. In 2009, my group and others showed how exome sequencing could be used to identify Mendelian-disease genes or to diagnose Mendelian disease[1,7,8]. Because there are tens of thousands of genetic variants in an exome, these strategies depended on effectively filtering out common variants, which are not likely to cause Mendelian disorders. At that time, databases of common variants were uneven and of suspect quality. Although ESP greatly improved the situation by uniformly and systematically cataloguing both common and rare variants across the exome[4], ExAC is an order of magnitude larger, and so enables better filtering. This is especially relevant for exome sequencing of non-European, non-African-American individuals, because ExAC provides greater sampling of individuals from outside the United States than ESP does.

On a related point, the study finds that hundreds of variants previously claimed to cause Mendelian disorders occur at implausibly high frequencies. As such, the authors suggest that they be reclassified as benign. A related study[9] shows how ExAC may also force a reassessment of whether some genes are involved at all in particular rare disorders. There is little doubt that ExAC will both refine and accelerate Mendelian-gene discovery and clinical genetics.

Finally, the consortium's approach to data aggregation and sharing is admirable. ExAC is both a technical and political achievement, requiring wrangling not only of data but also of investigators, consents and more from 14 studies — most of which were directed at the genetics of various common diseases.

An ongoing challenge in genomics is balancing the privacy rights of human participants with a strong tradition of promptly and openly sharing data. Building on the precedent of ESP, ExAC hits this balance by publicly releasing aggregate analyses —a catalogue of variants and the frequencies at which they arise — but not data about associated traits or other individual-level information (although raw data for many studies in ExAC is theoretically accessible through restricted databases). In this way, the study maximizes benefit while minimizing harm. These data have already been available on a terrifically intuitive website for nearly two years (http://exac.broadinstitute.org/), and the site has accrued more than 4 million page views.

If there is one take-home message, it is that there is incredible value in aggregating sequencing data across genomic studies. As the exomes aggregated by ExAC represent just a small fraction of the human samples that have been subjected to exome or genome sequencing so far, we can and should do better. In the coming decade, the number of human genomes that will be sequenced in some manner will grow to at least tens of millions and, by the end of this century, perhaps even billions. The beginnings of saturation seen here with CpG dinucleotides may eventually be observed deeply and at every site, providing a nucleotide-level footprint of the human genome. ■

Jay Shendure *is in the Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA, and is an investigator of the Howard Hughes Medical Institute.*
e-mail: shendure@uw.edu

1. Ng, S. B. *et al. Nature* **461,** 272–276 (2009).
2. Shendure, J. & Ji, H. *Nature Biotechnol.* **26,** 1135–1145 (2008).
3. Lek, M. *et al. Nature* **536,** 285–291 (2016).
4. Fu, W. *et al. Nature* **493,** 216–220 (2013).
5. The 1000 Genomes Project Consortium *Nature* **526,** 68–74 (2015).
6. The GTEx Consortium. *Science* **348,** 648–660 (2015).
7. Ng, S. B. *et al. Nature Genet.* **42,** 30–35 (2010).
8. Choi, M. *et al. Proc. Natl Acad. Sci. USA* **106,** 19096–19101 (2009).
9. Walsh, R. *et al. Genet. Med.* http://dx.doi.org/10.1038/GIM.2016.90 (2016).

NEUROSCIENCE

Flipping the sleep switch

Inactivation of a group of sleep-promoting neurons through dopamine signalling can cause acute or chronic wakefulness in flies, depending on changes in two different potassium-channel proteins. SEE LETTER P.333

STEPHANE DISSEL & PAUL J. SHAW

Many people have nodded off during a long road trip, or lain in bed desperately trying to fall asleep. These experiences illustrate real-world consequences of an improperly maintained balance between sleep- and wake-promoting neural circuits. On page 333, Pimentel *et al.*[1] describe the identification of a bona fide molecular switch that allows wake-promoting signals to turn off individual sleep-promoting neurons to regulate waking. These findings open up avenues for understanding the complexity of sleep regulation in healthy individuals and during disease.

Multiple sleep and wake circuits are found throughout the mammalian central nervous system and are believed to interact in a mutually inhibitory manner[2,3]. A similar organization is found in the fruitfly *Drosophila*, in which independent sleep and wake centres cooperate to produce stable sleep and wake patterns. Flies are less complex than mammals, and their neuronal circuits can be easily manipulated using genetic tools, making them more tractable as study subjects.

Perhaps the best-characterized sleep centre in flies is composed of neurons that project into a brain region called the dorsal fan-shaped body (dFB)[4–6]. The wake-promoting neurons (dopaminergic neurons) in the fly's brain release the neurotransmitter molecule dopamine. To better understand the molecular logic used by these two sets of neurons to regulate the sleep-to-wake transition, Pimentel *et al.* stimulated the wake-promoting neurons to release dopamine while they simultaneously recorded the activity of the sleep-promoting dFB neurons.

The authors genetically engineered the flies' dopaminergic neurons so that their activity could be modulated by a pulse of light — a technique known as optogenetics. Flies were restrained by fixing their heads, such that they could freely move their legs on a treadmill when awake. With this set-up, the physiological activity of the sleep-promoting dFB neurons and their response to the activation of dopaminergic wake-promoting neurons could be studied in real time.

When a fly spontaneously went to sleep for at least five minutes, the researchers optogenetically activated its dopaminergic neurons and measured the effect on the dFB neurons. In the sleeping flies, in the absence of dopamine signalling, the dFB neurons showed spikes of activity, which the researchers term the 'ON' state (Fig. 1a). Following acute activation of their dopaminergic neurons, the flies rapidly awoke. Pimentel and colleagues found that the dFB neurons were transiently hyperpolarized at this point (that is, the electrical potential across the cells' membranes sharply decreased), which inhibited neuronal firing. This change in membrane potential was brought about by a movement of potassium ions out of the cell, perhaps through a potassium-channel protein called Shaker that shows time- and voltage-dependent activation.

Next, Pimentel *et al.* applied dopamine directly to dFB dendrites — projections that receive signals from other neurons and transmit them to the body of the dFB neuron. Again, they observed hyperpolarization and suppression of firing, confirming that the effects of optogenetic activation reflected the direct consequences of increased dopamine signalling. The authors mapped this response to the Dop1R2 dopamine receptor protein, which is expressed in dFB neurons. Moreover, pharmacological and genetic studies demonstrated that Dop1R2 activation by dopamine inhibits dFB spiking through the $G_{i/o}$ signalling pathway, which in turn modulates physiological properties of the neuron. Thus, the researchers have found a molecular mechanism by which a wake-promoting dopamine signal directly and acutely inhibits dFB spiking during the ON state, allowing flies to rapidly wake up as environmental conditions demands (Fig. 1b).

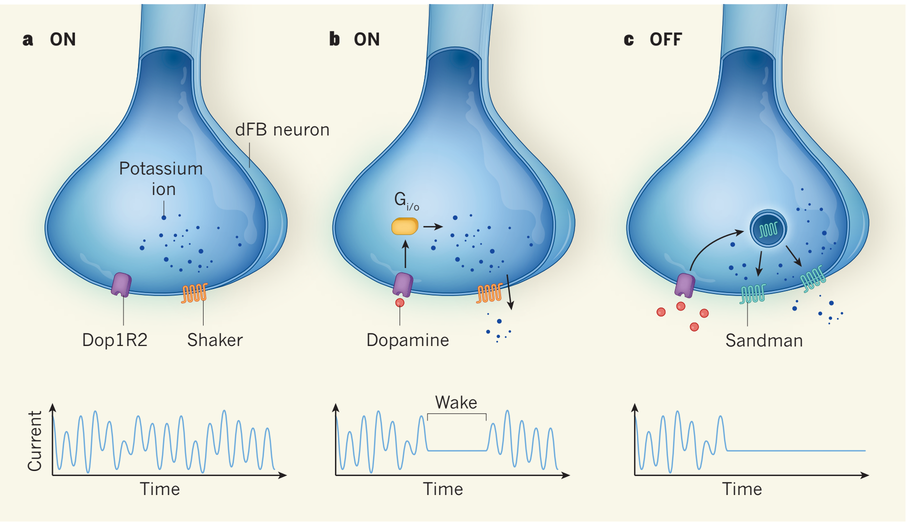


Figure 1 | Switch off to wake up. **a**, Pimentel *et al.*[1] report that, during sleep, neurons in the brain's dorsal fan-shaped body (dFB) in flies are in an ON state. A potassium-channel protein, probably Shaker, is closed, so potassium ions remain in the cell, and the dopamine receptor protein Dop1R2 is inactive. In this state, dFB neurons show repetitive bursts of activity known as spiking (displayed as peaks in a graph of electrical current). **b**, After transient dopamine release from wake-promoting neurons (not shown), the neurotransmitter molecule binds to and activates Dop1R2, triggering signalling through the protein $G_{i/o}$. This in turn opens and causes potassium efflux through Shaker, producing a current across the cell membrane that acutely inhibits spiking. Animals rapidly awaken while neurons remain in the ON state. **c**, Prolonged exposure to dopamine triggers an OFF state. Currents produced by Shaker are reduced (not shown), and another channel protein, Sandman, moves from vesicles in the cytoplasm to the membrane to mediate potassium efflux. This produces a different type of current that chronically inhibits neuronal firing.

In addition to this acute response, Pimentel and colleagues showed that longer exposures to dopamine turned off sleep-promoting neurons for extended periods of time. In this setting, dopamine hyperpolarized and inhibited spiking of dFB neurons through a different potassium channel, dubbed Sandman by the authors. During prolonged exposure to dopamine, Sandman channels, which are normally retained in the cytoplasm, are inserted into the membrane to chronically inhibit dFB firing. Following chronic inhibition, dFB neurons became impervious for up to an hour to a variety of stimuli that ordinarily trigger neuronal impulses . This 'OFF' state persisted even in the absence of a steady supply of dopamine, and produced a prolonged period of insomnia (Fig. 1c). Thus, the arousing effects of dopamine occur through two distinct mechanisms that operate on different timescales.

These results are truly exciting because they describe how circuit interactions are integrated at the level of individual neurons to maintain stable sleep–wake patterns. These types of analysis, particularly as they pertain to sleep, have proved to be inherently difficult in more-complex animals[7]. It will be interesting now to turn to mammals, looking at various types of sleep-promoting neuron to determine whether the switching mechanisms identified by Pimentel *et al.* in flies have similar roles in a more complex setting. In particular, it remains to be seen how this molecular-switching mechanism can be integrated into models that conceptualize sleep regulation as a winner-takes-all competition between opposing sleep-promoting and wake-promoting circuits[2].

Many avenues for further investigation remain. First, it will be interesting to find the switching mechanism that mediates the transition from waking to sleep, which has obvious clinical utility for conditions such as insomnia and Alzheimer's disease. Second, the dFB is made up of diverse populations of neurons that receive information from a large variety of neurotransmitters and neuropeptide molecules[8] — understanding how dopamine signalling interacts with these other neuroactive compounds will be crucial for truly understanding the molecular logic that underlies sleep regulation.

In summary, Pimentel and colleagues' study is valuable not only for the answers that it has given, but also because it provides the tools and logical framework that open up an area of enquiry that will be fruitful for many years to come. In particular, the ability to target a more precise molecular mechanism in discrete sets of neurons may make it easier to develop better drugs to enhance both sleep and waking with fewer adverse side effects. ■

Stephane Dissel *and* **Paul J. Shaw** *are in the Department of Neuroscience, Washington University School of Medicine in St. Louis,*

St. Louis, Missouri 63110, USA.
e-mails: shawp@wustl.edu; dissels@wustl.edu

1. Pimentel, D. *et al. Nature* **536,** 333–337 (2016).
2. Saper, C. B., Scammell, T. E. & Lu, J. *Nature* **437,** 1257–1263 (2005).
3. Jones, B. E. *Handb. Clin. Neurol.* **98,** 131–149 (2011).
4. Donlea, J. M., Thimgan, M. S., Suzuki, Y., Gottschalk, L. & Shaw, P. J. *Science* **332,** 1571–1576 (2011).
5. Liu, Q., Liu, S., Kodama, L., Driscoll, M. R. & Wu, M. N. *Curr. Biol.* **22,** 2114–2123 (2012).
6. Ueno, T. *et al. Nature Neurosci.* **15,** 1516–1523 (2012).
7. Sorooshyari, S., Huerta, R. & de Lecea, L. *Front. Neurol.* **6,** 32 (2015).
8. Kahsai, L. & Winther, A. M. *J. Comp. Neurol.* **519,** 290–315 (2011).

This article was published online on 3 August 2016.

CATALYSIS

Elusive active site in focus

The identification of the active site of an iron-containing catalyst raises hopes of designing practically useful catalysts for the room-temperature conversion of methane to methanol, a potential fuel for vehicles. SEE LETTER P.317

JAY A. LABINGER

On page 317, Snyder *et al.*[1] describe how they have attacked two of the most challenging problems in the field of catalysis using methods more common to the study of metalloenzymes. The first problem is how to pick out from an assembly of potential candidates the active site of a heterogeneous catalyst — a solid that can accelerate reactions of chemical species in the gas phase or in solution — and to determine its structure. The second problem is more specific: how to design an efficient process for selectively converting methane, the main component of natural gas, to a more valuable product. The authors' approach combines powerful spectroscopic techniques with computational modelling, and leads to a detailed picture of a catalytic site that is probably responsible for activating methane so that it can react at room temperature.

There is great interest in transforming methane into more useful liquid fuels such as methanol. But methane is notoriously unreactive, and most transformations require conditions, such as high temperatures, that are unfavourable for the selective formation of a desired product. One prominent exception occurs in microorganisms: certain bacteria use methane as a source of carbon and energy by first converting it to methanol, using enzymes known as methane monooxygenases (MMOs).

In 1997, an iron-containing structure that could be generated in certain zeolites was reported[2] to convert methane to methanol, even at room temperature (Fig. 1). (Zeolites are crystalline materials containing ordered arrays of pores that can house the active sites of catalysts.) One particularly intriguing aspect of this discovery was that soluble versions of MMO are also based on iron, and feature bimetallic ferrous oxide (Fe_2O_2) cores at their active site[3].

By using various spectroscopic techniques, several research groups have proposed[4,5] that the iron species of the zeolite system contains an analogous bimetallic structure, known as the α-Fe(II) centre. But such studies are complicated by the non-uniform nature of heterogeneous catalysts; it is difficult to determine whether a particular spectroscopic feature is associated with the actual active site, rather than reflecting an inactive 'spectator' site, or is a blurred-out average of both. From their interpretation of results using a suite of methods, Snyder *et al.* propose a quite different structure for the α-Fe(II) centre.

As a first step, the authors observed changes in the optical spectrum of the iron-containing zeolite as the material was cycled through the various stages of the methanol-forming reaction. This allowed them to assign particular peaks to the reactive states of the iron species α-Fe(II) and α-O, an intermediate species that forms during the reaction. They could then identify the corresponding features in the spectrum obtained using magnetic circular dichroism (MCD), a technique that is highly sensitive to the molecular and electronic structure of transition-metal centres such as iron. By comparing parameters determined by MCD — as well as the parameters' dependence on temperature and magnetic-field strength — with those of structurally well-characterized model compounds, the authors concluded that the most probable structure for α-Fe(II) contains a highly unusual, monometallic iron centre that has a 'square-planar' geometry (Fig. 1), whereas α-O is a square-pyramidal Fe(IV)=O structure resulting from the attachment of an oxygen ligand to the top of the square-planar species.

Snyder and colleagues' conclusions were further supported by Mössbauer spectroscopy — another technique often used in bioinorganic chemistry for structural characterization of iron centres — along with computational results. As in earlier work[4], Mössbauer spectroscopy on the iron-containing zeolite gave an overlay of several signals. This was potentially problematic, because it can be hard to find a priori grounds for assigning different Mössbauer signals to particular metal species. Fortunately, Snyder and co-workers were able to correlate the relative intensity of the largest signal quite precisely with the relative concentration of the active species, as determined from the amount of reaction product ultimately obtained. This provides more confidence that the authors have assigned the correct chemical structures to the active centres, and helps to demonstrate the potential of this multifaceted spectroscopic approach for characterizing heterogeneous catalysts.

There is one caveat: the system under study is particularly well suited to the authors' approach, for two reasons. First, the active site constitutes a large fraction of the total iron species in the zeolite, about 80% or more; and

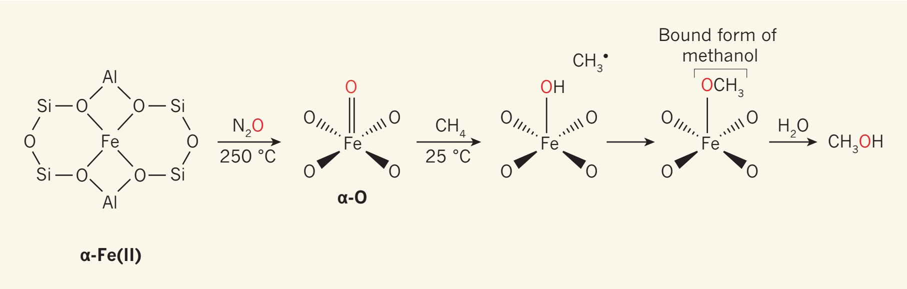


Figure 1 | Conversion of methane to methanol at room temperature. An iron-containing structure[2] called α-Fe(II) converts methane (CH_4) to methanol (CH_3OH) and can be generated in microporous crystalline materials called zeolites. Treatment of α-Fe(II) with nitrous oxide (N_2O) at 250 °C generates α-O, which contains a highly reactive form of oxygen (red atom, which originates from N_2O) that can readily remove a hydrogen atom from methane at 25 °C. The resulting methyl radical ($CH_3^\bullet$) combines with that oxygen to give a bound form of methanol, which can then be extracted from the zeolite by water. Snyder *et al.*[1] report that the iron atom in α-Fe(II) is constrained to an unusual 'square-planar' geometry within a ring of atoms in the zeolite; ring atoms are shown around the iron atom in α-Fe(II), but are not shown for the other iron-containing structures. Fe, iron; Si, silicon; Al, aluminium.

second, the reaction steps can be carried out one at a time because they occur under different conditions, allowing them to be correlated with the spectral changes they engender. One or both of those advantages probably will not apply in most heterogeneous catalytic systems of interest.

The proposed structure of α-Fe(ɪɪ) is intuitively pleasing to chemists because it makes sense for this species' unusual reactivity (cleavage of the strong carbon–hydrogen bond of methane at low temperatures) to be associated with an unusual structure (square-planar geometry is rare for Fe(ɪɪ) centres). The authors suggest that this structure is enforced by the rigid zeolite environment, in much the same way that proteins often constrain the active sites in metal-containing enzymes to abnormal geometries[6]. Furthermore, the fact that the structure is apparently quite different from that of the iron centre in MMOs, despite the similar reactivities, is an encouraging sign, because it suggests that various different iron species can solve the difficult problem of converting methane into methanol.

However, this iron-zeolite system is far from being a practical methane-conversion catalyst. Indeed, it is not really a catalyst at all: the various steps of the reaction each require very different conditions; and a complete reaction cycle, including release of the product from the zeolite, is completed only by using an extraction step, affording impractically low levels of methane conversion.

Furthermore, although the iron-containing zeolite successfully activates methane to undergo reactions, that is by no means the only prerequisite for an overall methane-conversion scheme, and often not even the most challenging one. Both thermodynamic and kinetic aspects of these reactions make it extremely difficult to oxidize methane selectively to methanol, without oxidizing it further and ultimately producing carbon dioxide[7]. In the zeolite, selectivity is achieved because the methanol is effectively immobilized by strong binding to the active site, which prevents it from undergoing over-oxidation at another zeolite site — but this also restricts methane conversion to extremely low levels. Operation of the system in a truly catalytic mode would expose the methanol to further oxidation, almost certainly decimating selectivity, as has been observed[8] when the zeolite is used at high temperatures (greater than 200 °C). Nonetheless, the insights gained by Snyder *et al.* provide information that may well aid the design of catalysts for the highly desirable conversion of methane to methanol. ■

Jay A. Labinger *is at the Beckman Institute, California Institute of Technology, Pasadena, California 91125, USA.*
e-mail: jal@caltech.edu

1. Snyder, B. E. R. *et al. Nature* **536,** 317–321 (2016).
2. Panov, G. I. *et al. React. Kinet. Catal. Lett.* **61,** 251–258 (1997).
3. Shu, L. *et al. Science* **275,** 515–518 (1997).
4. Dubkov, K. A., Ovanesyan, N. S., Shteinman, A. A., Starokon, E. V. & Panov, G. I. *J. Catal.* **207,** 341–352 (2002).
5. Xia, H. *et al. J. Phys. Chem. C* **112,** 9001–9005 (2008).
6. Vallee, B. L. & Williams, R. J. *Proc. Natl Acad. Sci. USA* **59,** 498–505 (1968).
7. Labinger, J. A. *J. Mol. Catal. A* **220,** 27–35 (2004).
8. Parfenov, M. V., Starokon, E. V., Pirutko, L. V. & Panov, G. I. *J. Catal.* **318,** 14–21 (2014).

MAMMALIAN DEVELOPMENT

Mechanics drives cell differentiation

Several hypotheses have been formulated to explain how cells make the first lineage decision during mammalian embryonic development. An overarching mechanism now unifies these disparate models. See Letter p.344

BERENIKA PLUSA & ANNA-KATERINA HADJANTONAKIS

The early mammalian embryo is an exemplar of a self-organizing system — distinct cell lineages are autonomously defined and stereotypically positioned as development progresses. The mechanism underlying formation of these cell lineages has long been elusive. On page 344, Maitre *et al.*[1] find that coordination between the contractility, polarity and position of a cell determines its identity, thereby defining the first lineage decision in the mouse embryo.

During the first stages of mammalian development, the fertilized egg undergoes a series of divisions that produce cells called blastomeres. During the transition from the 8- to the 16-cell stage, different cell lineages arise for the first time. Some blastomeres adopt an internal position and form the inner cell mass (ICM) from which the embryo proper will arise, whereas cells adopting an outer position become the trophectoderm layer[2,3] and go on to form the placenta.

Several models have been proposed for the regulation of this first cell-fate decision[4]. The first, put forward[5] in 1967 and later confirmed experimentally[6], posited that lineage is determined by the position of blastomeres within the embryo — whether or not they make contact with the external environment. When cell polarity emerged as a major feature of the lineage-specification process, an alternative mechanism was proposed[3]. At the 8-cell stage, blastomeres become polarized along their apical–basal axis, with certain proteins becoming restricted to the apical domain[2,7] (the side of the cell facing towards the outside of the embryo). This hypothesis stated that cells that inherit the

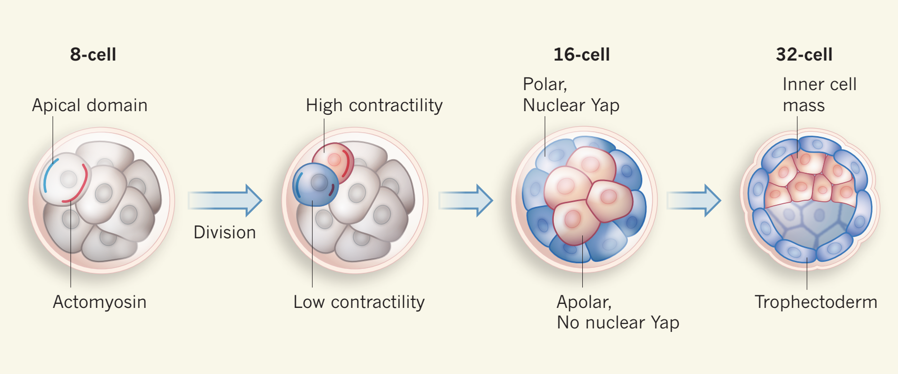


Figure 1 | A fateful decision. At the 8-cell stage of mouse development, the cells become polarized, with certain proteins becoming enriched on the apical side of the cell and forming an apical domain. As cells divide asymmetrically, one daughter inherits the apical domain and remains polarized (blue) and has low contractility, whereas the other inherits an abundance of the scaffold protein actomyosin and is apolar and highly contractile (red). Maitre *et al.*[1] report that these differences in contractility confer different fates at the 16-cell stage. In the less-contractile polar cells, the transcriptional activator protein Yap enters the nucleus and activates a gene-expression program that instructs the cell to become trophectoderm at the 32-cell stage, eventually giving rise to the placenta. Highly contractile cells do not have nuclear Yap, and adopt an inside position to become the inner cell mass, from which the embryo will form.

apical region at cell division acquire polarity, adopt an outside position and become trophectoderm, whereas those that do not inherit an apical region are internalized and become ICM.

The molecular mechanisms that link cell polarity to the first cell-fate choices remained a mystery for many years. But recently, differential activity of the Hippo signalling pathway was shown to be crucial for the decision[8]. Despite this advance, exactly how position and polarity cues translate into differences in Hippo pathway activity remained unclear. Adding to this disparate body of knowledge, accumulating evidence[9,10] indicated that a cell's position depends on its contractility. Maitre and colleagues' study[1] combines theory and experiment to unify the existing models of ICM–trophectoderm fate choice, and also provides a mechanistic link between cell polarity, position and Hippo pathway activity (Fig. 1).

The authors showed that asymmetric segregation of a polarized apical domain at cell division generates two daughter blastomeres with differential levels of contractility. Daughter cells that receive the apical domain are less contractile than their apolar sisters. The researchers then found that apolar blastomeres have higher levels of the scaffolding protein actomyosin than their polarized counterparts, directly translating into increased contractility. These differences in contractility trigger the sorting of cells to internal or external positions, because the less-contractile polarized cells have a tendency to spread over the apolar cells, which become internalized.

Support for this model came from a series of experiments in which Maitre *et al.* measured the surface tension of individual blastomeres, and then traced those cells over time in embryos to determine which lineage they adopted. This involved the development of technically sophisticated methods to probe the mechanics of individual cells and then track cells in embryos. Furthermore, the authors found that altering a cell's contractility altered its fate.

Finally, they demonstrated that contractility controls the subcellular position of the transcriptional co-activator protein Yap, a central component of the Hippo pathway. In less-contractile, polarized cells, Yap translocated to the nucleus, leading to activation of trophectoderm-specific genes. In apolar, highly contractile cells, Yap remained excluded from the nucleus. Linking Yap activity to differences in cell contractility connects the mechanical properties of blastomeres to their cell-fate choices, suggesting that mechanosensing may affect early lineage decisions.

Mammalian embryos are renowned for their ability to develop normally following alterations in internal architecture, or the loss or addition of cells. It has been proposed[11] that activation of dormant mechanisms might help embryos to successfully carry out development following perturbations. Indeed, Maitre *et al.* showed that the mechanism responsible for ICM or trophectoderm specification is also probably used to compensate for perturbations, and thus underpins the regulative nature of mammalian embryos. By mixing cells with differential contractility, the authors demonstrated that those with elevated contractility adopted an internal position within embryos, whereas those with reduced contractility adopted an outside position. A mechanistic link between the position of a cell, its contractility and its gene-expression profile explains how the cell might 'sense' and consequently 'adjust' its position in the embryo, altering gene expression accordingly.

Although the current study represents a major breakthrough in our understanding of early mammalian development, several questions remain open. For instance, it is still not clear what triggers the initial blastomere polarization and differential contractility. It has been shown that modulating key transcription factors that control cell fate can influence cell position within an embryo, and so a feedback mechanism perhaps translates changes in gene expression into changes in contractility. In addition, the mechanism by which actomyosin affects the subcellular positioning of Yap needs to be determined.

Perhaps most important is that the results of Maitre and colleagues' study beg the question of whether the same mechanism is used in a variety of developmental contexts. Is mechanosensing through cellular contractility repeatedly used to regulate a cell's propensity to adopt alternative fates? The answer is sure to provide valuable insights into how lineage decisions are made in the mammalian embryo. ■

Berenika Plusa *is in the Faculty of Life Sciences, University of Manchester, Manchester M13 9PT, UK.* **Anna-Katerina Hadjantonakis** *is in the Developmental Biology Program, Sloan Kettering Institute, Memorial Sloan Kettering Cancer Center, New York, New York 10065, USA.*
e-mails: berenika.plusa@manchester.ac.uk; hadj@mskcc.org

1. Maitre, J.-L. *et al.* Nature **536,** 344–348 (2016).
2. Johnson, M. H. & McConnell, J. M. *Semin. Cell Dev. Biol.* **15,** 583–597 (2004).
3. Johnson, M. H. & Ziomek, C. A. *Cell* **24,** 71–80 (1981).
4. Chazaud, C. & Yamanaka, Y. *Development* **143,** 1063–1074 (2016).
5. Tarkowski, A. K. & Wróblewska, J. *J. Embryol. Exp. Morphol.* **18,** 155–180 (1967).
6. Hillman, N., Sherman, M. I. & Graham, C. *J. Embryol. Exp. Morphol.* **28,** 263–278 (1972).
7. Plusa, B. *et al. J. Cell Sci.* **118**, 505–515 (2005).
8. Sasaki, H. *Semin. Cell Dev. Biol.* **47–48,** 80–87 (2015).
9. Anani, S., Bhat, S., Honma-Yamanaka, N., Krawchuk, D. & Yamanaka, Y. *Development* **141,** 2813–2824 (2014).
10. Samarage, C. R. *et al. Dev. Cell* **34,** 435–447 (2015).
11. Piotrowska, K. & Zernicka-Goetz, M. *Nature* **409,** 517–521 (2001).

This article was published online on 3 August 2016.

CONDENSED-MATTER PHYSICS

Superconducting electrons go missing

'Overdoped' high-temperature superconductors, which have a high density of charge carriers, were thought to be well understood. An experiment challenges what we know about quantum physics in such systems. SEE LETTER P.309

JAN ZAANEN

Put a large number of interacting quantum particles together and exotic things can happen. A landmark example is superconductivity[1], in which certain materials have zero electrical resistance when cooled below a critical temperature, T_c. The first such materials to be discovered were superconductive only at low temperatures (T_c of a few kelvin), and their behaviour could be explained by the Bardeen–Cooper–Schrieffer (BCS) theory[2]. Superconductivity was considered a closed chapter until the surprising discovery in 1986 of high-temperature copper oxide superconductors[3] (T_c of up to 160 K), whose properties couldn't be described by BCS theory[4]. Nevertheless, it was thought that overdoping such superconductors — significantly increasing the density of charge carriers and thereby reducing the materials' T_c — would bring them into agreement with the theory[4]. Božović *et al.*[5] show on page 309 that even this overdoped regime is highly anomalous, a finding that has implications for our fundamental understanding of superconductivity.

In quantum statistics, particles called bosons can exist in the same quantum state, whereas fermion particles must occupy different states — a restriction called the Pauli exclusion principle. If many bosons occupy the lowest energy state of a system (a configuration known as a Bose condensate), the microscopic quantum behaviour gets amplified to the macroscopic scale, and the result is superconductivity.

BCS theory explains how electrons can form a Bose condensate, even though they are fermions. The natural state of fermions is as a Fermi gas, in which the particles fill up the lowest energy levels of a system, according to the Pauli exclusion principle. The boundary between the filled and unfilled energy levels is known as the Fermi surface. In BCS theory, upon introducing a small attractive interaction between the electrons in a system, those electrons closest to the Fermi surface bind together to form what are called Cooper pairs[1,2]. Because these Cooper pairs are effectively bosons, they form a Bose condensate.

One might expect that only the small fraction of electrons that form Cooper pairs contributes to superconductivity, but in fact all the electrons in the Fermi gas participate. The density of superconductive electrons (the superfluid density) is therefore approximately equal to the total electron density — a prediction that has been confirmed experimentally in many 'conventional' superconductors[1].

The situation is less straightforward for copper oxide superconductors because the electrons in these systems interact strongly[4]. The electrons behave like cars on a motorway[6], forming a traffic jam (a 'Mott insulator'[7]) when the density is high. Doping the system is equivalent to reducing the density of cars on the motorway, which gives rise to a kind of stop–start traffic. The result is that T_c increases with doping, reaching a maximum at a level known as optimal doping.

In the overdoped regime (when doping has increased beyond optimal doping), the cars (electrons) move more freely and no longer interact strongly, leading to a reduction in T_c. In this weakly interacting regime, one might expect the system to be described by BCS theory[4]. This expectation seemed to be confirmed when a Fermi surface subjected to textbook Cooper pairing was directly observed in a variety of overdoped systems[8,9]. Because, in BCS theory, the superfluid density is approximately equal to the total electron density, it should be large and almost independent of both doping and T_c.

Božović and colleagues present the first reliable measurements of the superfluid density in the overdoped regime. Such measurements have taken so long to obtain because the material is difficult to prepare: overdoped copper oxides are chemically unstable. But, in an impressive feat of materials engineering, the authors manage to prepare near-perfect samples using sophisticated techniques.

By measuring the superfluid density as a function of doping, the authors find that there are far fewer superconducting electrons than expected from BCS theory (Fig. 1a) — most of these electrons seem to be missing. The authors also demonstrate a simple scaling law (let's call it Božović's scaling law): the superfluid density is directly proportional to T_c over the entire overdoping range (Fig. 1b). Given the experimental evidence suggesting that BCS theory is at work in this regime[8,9], the findings present a paradox — Božović's scaling law disagrees fundamentally with the predictions of BCS theory, and therefore comes as a complete surprise.

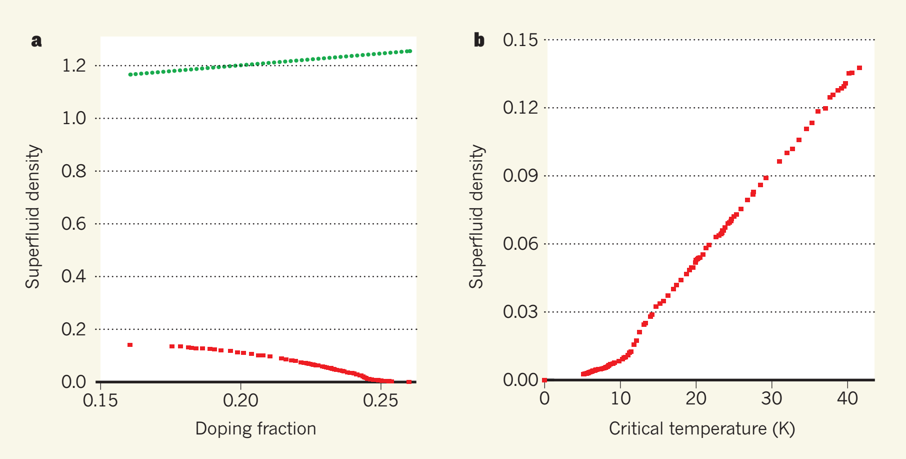


Figure 1 | Surprising superconductivity. Božović *et al.*[5] have measured key properties of 'overdoped' high-temperature copper oxide superconductors. **a**, From their measurements, the authors estimate the superfluid density (red, corresponding to the number of electrons taking part in superconductivity per unit cell; the unit cell is the smallest periodically repeating structure in a crystal) as a function of the doping fraction (a measure of the density of charge carriers). They observe far fewer superconducting electrons than expected from the Bardeen–Cooper–Schrieffer (BCS) theory (green). **b**, The data are presented here as a plot of the superfluid density against the critical temperature (the temperature below which the material can superconduct). The superfluid density is directly proportional to the critical temperature, over a wide doping range. The authors' remarkable results are incompatible with standard BCS theory and require new explanations.

This paradox is partly resolved by a theoretical loophole. Leggett's theorem[10] describes the conditions that must be met for the superfluid density of a system to be equal to the total electron density. One condition is that the system must be translationally invariant — at the atomic scale, the system must be the same at each point in space.

In superconductors, translational invariance does not occur at the atomic scale because the electron system exists in a periodic lattice that is formed from atoms. But in conventional superconductors, quantum physics causes low-energy electronic excitations to behave as though the lattice isn't there. Translational invariance is therefore realized as an emergent symmetry (a symmetry that is seen only on large scales), which means that a BCS superconductor must obey Leggett's theorem.

However, for high-temperature copper oxide superconductors, Božović and colleagues' results suggest that the electron system as a whole remains strongly interacting, even in the overdoped regime. Although such strongly interacting systems are poorly understood, there is experimental evidence that, in the underdoped regime, the superfluid density is not governed by Leggett's theorem[11] because the electrons are greatly affected by the presence of the lattice. Therefore, high-temperature superconductors might not obey Leggett's theorem in the overdoped regime either.

Paradoxes are extremely useful in science: the simple relationship between the superfluid density and T_c suggests that some underlying principle must be at work in overdoped copper oxide superconductors. A similar scaling law has been observed[11] in the underdoped regime, and is explained in terms of electrons binding together in pairs at high temperatures and forming a Bose condensate at T_c. However, this explanation cannot apply in the overdoped regime because of the observed Fermi surfaces. Indeed, there is nothing in the vast literature of superconductivity research that sheds light on this conundrum, and Božović's scaling law forces physicists to go back to the drawing board. ■

Jan Zaanen *is at the Instituut-Lorentz for Theoretical Physics, Leiden University, 2300 RA Leiden, the Netherlands.*
e-mail: jan@lorentz.leidenuniv.nl

1. Tinkham, M. *Introduction to Superconductivity* (Dover, 2004).
2. Bardeen, J., Cooper, L. N. & Schrieffer, J. R. *Phys. Rev.* **106,** 162–164 (1957).
3. Bednorz, J. G. & Müller, K. A. *Z. Phys. B* **64,** 189–193 (1986).
4. Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. *Nature* **518,** 179–186 (2015).
5. Božović, I., He, X., Wu, J. & Bollinger, A. T. *Nature* **536,** 309–311 (2016).
6. Zaanen, J. *Science* **315,** 1372–1373 (2007).
7. Mott, N. F. *Proc. Phys. Soc. A* **62,** 416–422 (1949).
8. Vignolle, B. *et al. Nature* **455,** 952–955 (2008).
9. Chatterjee, U. *et al. Proc. Natl Acad. Sci. USA* **108,** 9346–9349 (2011).
10. Leggett, A. J. *J. Stat. Phys.* **93,** 927–941 (1998).
11. Uemura, Y. J. *et al. Phys. Rev. Lett.* **62,** 2317–2320 (1989).

OPEN
doi:10.1038/nature19057

Analysis of protein-coding genetic variation in 60,706 humans

Monkol Lek[1,2,3,4], Konrad J. Karczewski[1,2]*, Eric V. Minikel[1,2,5]*, Kaitlin E. Samocha[1,2,5,6]*, Eric Banks[2], Timothy Fennell[2], Anne H. O'Donnell-Luria[1,2,7], James S. Ware[2,8,9,10,11], Andrew J. Hill[1,2,12], Beryl B. Cummings[1,2,5], Taru Tukiainen[1,2], Daniel P. Birnbaum[2], Jack A. Kosmicki[1,2,6,13], Laramie E. Duncan[1,2,6], Karol Estrada[1,2], Fengmei Zhao[1,2], James Zou[2], Emma Pierce-Hoffman[1,2], Joanne Berghout[14,15], David N. Cooper[16], Nicole Deflaux[17], Mark DePristo[18], Ron Do[19,20,21,22], Jason Flannick[2,23], Menachem Fromer[1,6,19,20,24], Laura Gauthier[18], Jackie Goldstein[1,2,6], Namrata Gupta[2], Daniel Howrigan[1,2,6], Adam Kiezun[18], Mitja I. Kurki[2,25], Ami Levy Moonshine[18], Pradeep Natarajan[2,26,27,28], Lorena Orozco[29], Gina M. Peloso[2,27,28], Ryan Poplin[18], Manuel A. Rivas[2], Valentin Ruano-Rubio[18], Samuel A. Rose[6], Douglas M. Ruderfer[19,20,24], Khalid Shakir[18], Peter D. Stenson[16], Christine Stevens[2], Brett P. Thomas[1,2], Grace Tiao[18], Maria T. Tusie-Luna[30], Ben Weisburd[2], Hong-Hee Won[31], Dongmei Yu[6,25,27,32], David M. Altshuler[2,33], Diego Ardissino[34], Michael Boehnke[35], John Danesh[36], Stacey Donnelly[2], Roberto Elosua[37], Jose C. Florez[2,26,27], Stacey B. Gabriel[2], Gad Getz[18,26,38], Stephen J. Glatt[39,40,41], Christina M. Hultman[42], Sekar Kathiresan[2,26,27,28], Markku Laakso[43], Steven McCarroll[6,8], Mark I. McCarthy[44,45,46], Dermot McGovern[47], Ruth McPherson[48], Benjamin M. Neale[1,2,6], Aarno Palotie[1,2,5,49], Shaun M. Purcell[19,20,24], Danish Saleheen[50,51,52], Jeremiah M. Scharf[2,6,25,27,32], Pamela Sklar[19,20,24,53,54], Patrick F. Sullivan[55,56], Jaakko Tuomilehto[57], Ming T. Tsuang[58], Hugh C. Watkins[44,59], James G. Wilson[60], Mark J. Daly[1,2,6], Daniel G. MacArthur[1,2] & Exome Aggregation Consortium†

Large-scale reference data sets of human genetic variation are critical for the medical and functional interpretation of DNA sequence changes. Here we describe the aggregation and analysis of high-quality exome (protein-coding region) DNA sequence data for 60,706 individuals of diverse ancestries generated as part of the Exome Aggregation Consortium (ExAC). This catalogue of human genetic diversity contains an average of one variant every eight bases of the exome, and provides direct evidence for the presence of widespread mutational recurrence. We have used this catalogue to calculate objective metrics of pathogenicity for sequence variants, and to identify genes subject to strong selection against various classes of mutation; identifying 3,230 genes with near-complete depletion of predicted protein-truncating variants, with 72% of these genes having no currently established human disease phenotype. Finally, we demonstrate that these data can be used for the efficient filtering of candidate disease-causing variants, and for the discovery of human 'knockout' variants in protein-coding genes.

[1]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [3]School of Paediatrics and Child Health, University of Sydney, Sydney, New South Wales 2145, Australia. [4]Institute for Neuroscience and Muscle Research, Children's Hospital at Westmead, Sydney, New South Wales 2145, Australia. [5]Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, Massachusetts 02115, USA. [6]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [7]Division of Genetics and Genomics, Boston Children's Hospital, Boston, Massachusetts 02115, USA. [8]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [9]National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK. [10]NIHR Royal Brompton Cardiovascular Biomedical Research Unit, Royal Brompton Hospital, London SW3 6NP, UK. [11]MRC Clinical Sciences Centre, Imperial College London, London SW7 2AZ, UK. [12]Genome Sciences, University of Washington, Seattle, Washington 98195, USA. [13]Program in Bioinformatics and Integrative Genomics, Harvard Medical School, Boston, Massachusetts 02115, USA. [14]Mouse Genome Informatics, Jackson Laboratory, Bar Harbor, Maine 04609, USA. [15]Center for Biomedical Informatics and Biostatistics, University of Arizona, Tucson, Arizona 85721, USA. [16]Institute of Medical Genetics, Cardiff University, Cardiff CF10 3XQ, UK. [17]Google, Mountain View, California 94043, USA. [18]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [19]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [20]Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [21]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [22]The Center for Statistical Genetics, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [23]Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [24]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [25]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [26]Harvard Medical School, Boston, Massachusetts 02115, USA. [27]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [28]Cardiovascular Research Center, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [29]Immunogenomics and Metabolic Disease Laboratory, Instituto Nacional de Medicina Genómica, Mexico City 14610, Mexico. [30]Molecular Biology and Genomic Medicine Unit, Instituto Nacional de Ciencias Médicas y Nutrición, Mexico City 14080, Mexico. [31]Samsung Advanced Institute for Health Sciences and Technology (SAIHST), Sungkyunkwan University, Samsung Medical Center, Seoul, South Korea. [32]Department of Neurology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [33]Vertex Pharmaceuticals, Boston, Massachusetts 02210, USA. [34]Department of Cardiology, University Hospital, 43100 Parma, Italy. [35]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan 48109, USA. [36]Department of Public Health and Primary Care, Strangeways Research Laboratory, Cambridge CB1 8RN, UK. [37]Cardiovascular Epidemiology and Genetics, Hospital del Mar Medical Research Institute, 08003 Barcelona, Spain. [38]Department of Pathology and Cancer Center, Massachusetts General Hospital, Boston, Massachusetts, 02114 USA. [39]Psychiatric Genetic Epidemiology & Neurobiology Laboratory, State University of New York, Upstate Medical University, Syracuse, New York 13210, USA. [40]Department of Psychiatry and Behavioral Sciences, State University of New York, Upstate Medical University, Syracuse, New York 13210, USA. [41]Department of Neuroscience and Physiology, State University of New York, Upstate Medical University, Syracuse, New York 13210, USA. [42]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, SE-171 77 Stockholm, Sweden. [43]Department of Medicine, University of Eastern Finland and Kuopio University Hospital, 70211 Kuopio, Finland. [44]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX1 2JD, UK. [45]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford OX1 2JD, UK. [46]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals Foundation Trust, Oxford OX1 2JD, UK. [47]Inflammatory Bowel Disease and Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA. [48]Atherogenomics Laboratory, University of Ottawa Heart Institute, Ottawa, Ontario K1Y 4W7, Canada. [49]Institute for Molecular Medicine Finland (FIMM), University of Helsinki, 00100 Helsinki, Finland. [50]Department of Biostatistics and Epidemiology, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [51]Department of Medicine, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA. [52]Center for Non-Communicable Diseases, Karachi, Pakistan. [53]Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [54]Department of Neuroscience, Icahn School of Medicine at Mount Sinai, New York, New York 10029, USA. [55]Department of Genetics, University of North Carolina, Chapel Hill, North Carolina 27599, USA. [56]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet SE-171 77 Stockholm, Sweden. [57]Department of Public Health, University of Helsinki, 00100 Helsinki, Finland. [58]Department of Psychiatry, University of California, San Diego, California 92093, USA. [59]Radcliffe Department of Medicine, University of Oxford, Oxford OX1 2JD, UK. [60]Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, Mississippi 39216, USA.

†A list of participants and their affiliations appears in the Supplementary Information.

*These authors contributed equally to this work.

Over the last five years, the widespread availability of high-throughput DNA sequencing technologies has permitted the sequencing of the whole genomes or exomes of hundreds of thousands of humans. In theory, these data represent a powerful source of information about the global patterns of human genetic variation, but in practice, are difficult to access for practical, logistical, and ethical reasons; in addition, their utility is complicated by the heterogeneity in the experimental methodologies and variant calling pipelines used to generate them. Current publicly available data sets of human DNA sequence variation contain only a small fraction of all sequenced samples: the Exome Variant Server, created as part of the NHLBI Exome Sequencing Project (ESP)[1], contains frequency information spanning 6,503 exomes; and the 1000 Genomes Project (1000G), which includes individual-level genotype data from whole-genome and exome sequence data for 2,504 individuals[2].

Databases of genetic variation are important for our understanding of human population history and biology[1–5], but also provide critical resources for the clinical interpretation of variants observed in patients who have rare Mendelian diseases[6,7]. The filtering of candidate variants by frequency in unselected individuals is a key step in any pipeline for the discovery of causal variants in Mendelian disease patients, and the efficacy of such filtering depends on both the size and the ancestral diversity of the available reference data.

Here we describe the joint variant calling and analysis of high-quality variant calls across 60,706 human exomes, assembled by the Exome Aggregation Consortium (ExAC; http://exac.broadinstitute.org). This call set exceeds previously available exome-wide variant databases, by nearly an order of magnitude, providing substantially increased resolution for the analysis of very low-frequency genetic variants. We demonstrate the application of this data set to the analysis of patterns of genetic variation including the discovery of widespread mutational recurrence, the inference of gene-level constraint against truncating variation, the clinical interpretation of variation in Mendelian disease genes, and the discovery of human knockout variants in protein-coding genes.

The ExAC data set

Sequencing data processing, variant calling, quality control and filtering was performed on over 91,000 exomes (see Methods), and sample filtering was performed to produce a final data set spanning 60,706 individuals (Fig. 1a). To identify the ancestry of each ExAC individual, we performed principal component analysis (PCA) to distinguish the major axes of geographic ancestry and to identify population clusters corresponding to individuals of European, African, South Asian, East Asian, and admixed American (hereafter referred to as Latino) ancestry (Fig. 1b; Supplementary Table 3); we note that the apparent separation between East Asian and other samples reflects a deficiency of Middle Eastern and Central Asian samples in the data set. We further separated Europeans into individuals of Finnish and non-Finnish ancestry given the enrichment of this bottlenecked population; the term European hereafter refers to non-Finnish European individuals.

We identified 10,195,872 candidate sequence variants in ExAC. We further applied stringent depth and site/genotype quality filters to define a subset of 7,404,909 high-quality variants, including 317,381 insertions or deletions (indels) (Supplementary Table 7), corresponding to one variant for every 8 base pairs (bp) within the exome intervals. The majority of these are very low-frequency variants absent from previous smaller call sets (Fig. 1c), of the high-quality variants, 99% have a frequency of <1%, 54% are singletons (variants seen only once in the data set), and 72% are absent from both 1000G and ESP data sets.

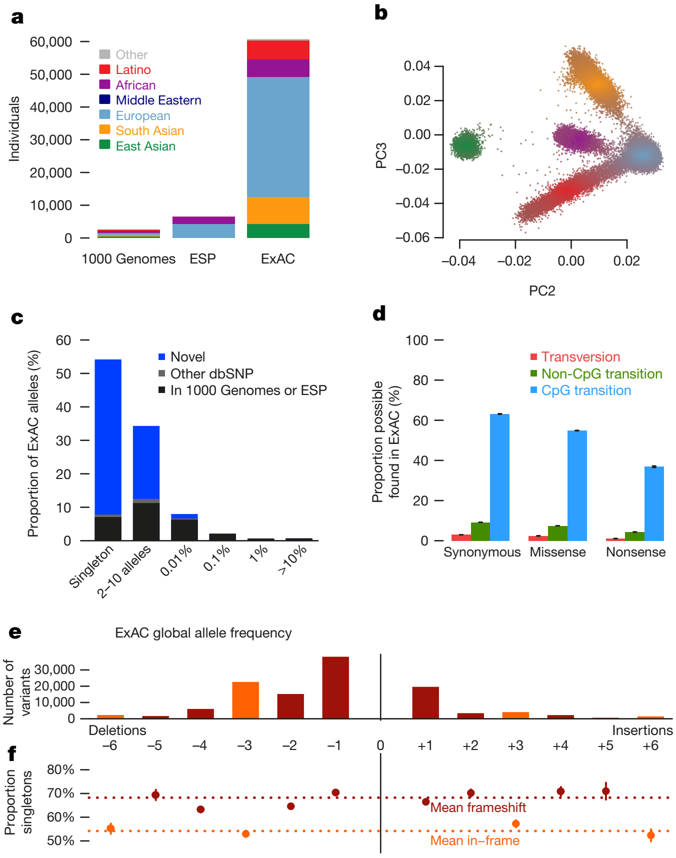


Figure 1 | Patterns of genetic variation in 60,706 humans. a, The size and diversity of public reference exome data sets. ExAC exceeds previous data sets in size for all studied populations. **b**, Principal component analysis (PCA) dividing ExAC individuals into five continental populations. PC2 and PC3 are shown; additional PCs are in Extended Data Fig. 5a. **c**, The allele frequency spectrum of ExAC highlights that the majority of genetic variants are rare and novel (absent from prior databases of genetic variation, such as dbSNP). **d**, The proportion of possible variation observed by mutational context and functional class. Over half of all possible CpG transitions are observed. Error bars represent standard error of the mean. **e**, **f**, The number (**e**), and frequency distribution (proportion singleton; **f**) of indels, by size. Compared to in-frame indels, frameshift variants are less common (have a higher proportion of singletons, a proxy for predicted deleteriousness on gene product). Error bars indicate 95% confidence intervals.

The density of variation in ExAC is not uniform across the genome, and the observation of variants depends on factors such as mutational properties and selective pressures. In the ~45 million well-covered (80% of individuals with a minimum of 10× coverage) positions in ExAC, there are ~18 million possible synonymous variants, of which we observe 1.4 million (7.5%). However, we observe 63.1% of possible CpG transitions (C to T variants, in which the adjacent base is G), while only observing 3% of possible transversions and 9.2% of other possible transitions (Supplementary Table 9). A similar pattern is observed for missense and nonsense variants, with lower proportions due to selective pressures (Fig. 1d). Of 123,629 high-quality indels called in coding exons, 117,242 (95%) have a length <6 bases, with shorter deletions being the most common (Fig. 1e). Frameshifts are found in smaller numbers and are more likely to be singletons than in-frame indels (Fig. 1f), reflecting the influence of purifying selection.

Patterns of protein-coding variation

The density of protein-coding sequence variation in ExAC reveals a number of properties of human genetic variation that are undetectable in smaller data sets. For example, 7.9% of high-quality sites in ExAC are multiallelic (multiple different sequence variants observed at the same site), close to the Poisson expectation of 8.3%, given the observed density of variation, and far higher than that observed in previous data sets of 0.48% in the 1000G (exome intervals) and 0.43% in the ESP data sets.

The size of ExAC makes it possible to directly observe mutational recurrence: instances in which the same mutation has occurred multiple times independently throughout the history of the sequenced populations. For instance, among synonymous (non-protein-altering) variants, a class of variation expected to have undergone minimal selection, 43% of validated *de novo* events identified in external data sets of 1,756 parent-offspring trios[8,9] are also observed independently in our data set (Fig. 2a), indicating a separate origin for the same variant within the demographic history of the two samples. This proportion is much higher for transition variants at CpG sites, well established to be the most highly mutable sites in the human genome[10]: 87% of previously reported *de novo* CpG transitions at synonymous sites are observed in ExAC, indicating that our sample sizes are beginning to approach saturation of this class of variation. This saturation is detectable by a change in the discovery rate at subsets of the ExAC data set, beginning at around 20,000 individuals (Fig. 2b), indicating that ExAC is the first human exome-wide data set, to our knowledge, large enough for this effect to be directly observed.

Mutational recurrence has a marked effect on the frequency spectrum in the ExAC data, resulting in a depletion of singletons at sites with high mutation rates (Fig. 2c). We observe a correlation between singleton rates (the proportion of variants seen only once in ExAC) and site mutability inferred from sequence context[11] ($r = -0.98$; $P < 10^{-50}$; Extended Data Fig. 1d): sites with low predicted mutability have a singleton rate of 60%, compared to 20% for sites with the highest predicted rate (CpG transitions; Fig. 2c). Conversely, for synonymous variants, CpG variants are approximately twice as likely to rise to intermediate frequencies: 16% of CpG variants are found in at least 20 copies in ExAC, compared to 8% of transversions and non-CpG transitions, suggesting that synonymous CpG transitions have on average two independent mutational origins in the ExAC sample. Recurrence at highly mutable sites can further be observed by examining the population sharing of doubleton synonymous variants (variants occurring in only two individuals in ExAC). Low-mutability mutations (especially transversions), are more likely to be observed in a single population (representing a single mutational origin), whereas CpG transitions are more likely to be found in two separate populations (independent mutational events); as such, site mutability and probability of observation in two populations is significantly correlated ($r = 0.884$; Fig. 2d).

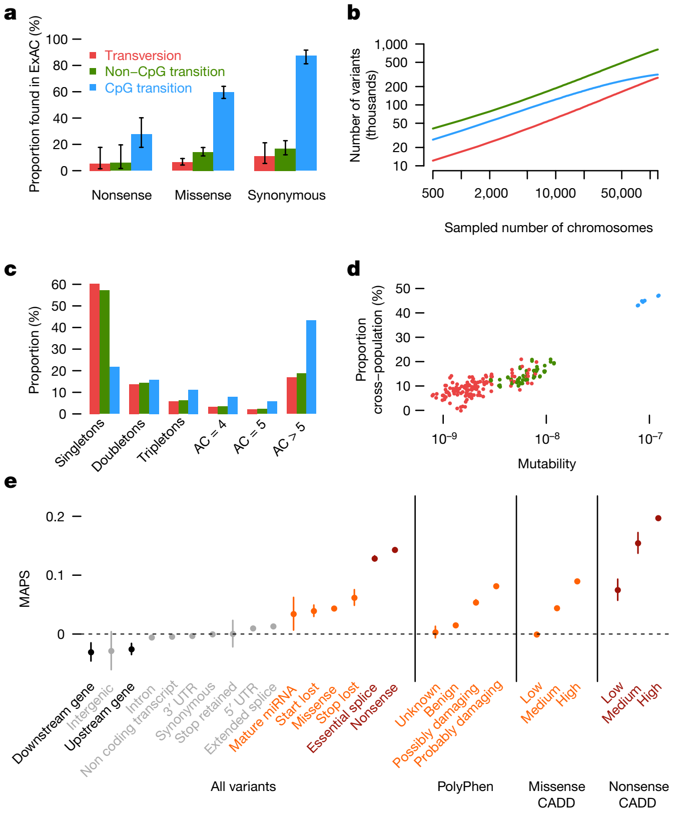


Figure 2 | Mutational recurrence at large sample sizes. a, Proportion of validated *de novo* variants from two external data sets that are independently found in ExAC, separated by functional class and mutational context. Error bars represent standard error of the mean. Colours are consistent in **a**–**d**. **b**, Number of unique variants observed, by mutational context, as a function of number of individuals (downsampled from ExAC). CpG transitions, the most likely mutational event, begin reaching saturation at ∼20,000 individuals. **c**, The site frequency spectrum is shown for each mutational context. **d**, For doubletons (variants with an allele count (AC) of 2), mutation rate is positively correlated with the likelihood of being found in two individuals of different continental populations. **e**, The mutability-adjusted proportion of singletons (MAPS) is shown across functional classes. Error bars represent standard error of the mean of the proportion of singletons.

We also explored the prevalence and functional impact of multinucleotide polymorphisms (MNPs), in cases where multiple substitutions were observed within the same codon in at least one individual. We found 5,945 MNPs (mean = 23 per sample) in ExAC (Extended Data Fig. 2a), in which analysis of the underlying SNPs without correct haplotype phasing would result in altered interpretation. These include 647 instances in which the effect of a protein-truncating variant (PTV) is eliminated by an adjacent single nucleotide polymorphism (SNP) (referred to as a rescued PTV), and 131 instances in which underlying synonymous or missense variants result in PTV MNPs (referred to as a gained PTV). Our analysis also revealed 8 MNPs in disease-associated genes, resulting in either a rescued or gained PTV, and 10 MNPs that have previously been reported as disease-causing mutations (Supplementary Tables 10 and 11). These variants would be missed by virtually all currently available variant calling and annotation pipelines.

Inferring variant deleteriousness and gene constraint

Deleterious variants are expected to have lower allele frequencies than neutral ones, due to negative selection. This theoretical property has been demonstrated previously in human population sequencing data[12,13] and here (Fig. 1d, e). This allows inference of the degree of selection against specific functional classes of variation. However, mutational recurrence as described earlier indicates that allele frequencies observed in ExAC-scale samples are also skewed by mutation rate, with more mutable sites less likely to be singletons (Fig. 2c and Extended Data Fig. 1d). Mutation rate is in turn non-uniformly distributed across functional classes. For example, variants that result in the loss of a stop codon can never occur at CpG dinucleotides (Extended Data Fig. 1e). We corrected for mutation rates (Supplementary Information section 3.2) by creating a mutability-adjusted proportion singleton (MAPS) metric. This metric reflects (as expected), strong selection against predicted PTVs, as well as missense variants predicted by conservation-based methods to be deleterious (Fig. 2e).

The deep ascertainment of rare variation in ExAC also allows us to infer the extent of selection against variant categories on a per-gene basis by examining the proportion of variation that is missing compared to expectations under random mutation. Conceptually similar approaches have been applied to smaller exome data sets[11,14], but have been underpowered, particularly when analysing the depletion of PTVs. We compared the observed number of rare (minor allele frequency (MAF) <0.1%) variants per gene to an expected number derived from a selection neutral, sequence-context based mutational model[11]. The model performs well in predicting the number of synonymous variants, which should be under minimal selection, per gene ($r = 0.98$; Extended Data Fig. 3b).

We quantified deviation from expectation with a *Z* score[11], which for synonymous variants is centred at zero, but is significantly shifted towards higher values (greater constraint) for both missense and PTV (Wilcoxon $P < 10^{-50}$ for both; Fig. 3a). The genes on the X chromosome are significantly more constrained than those on the autosomes

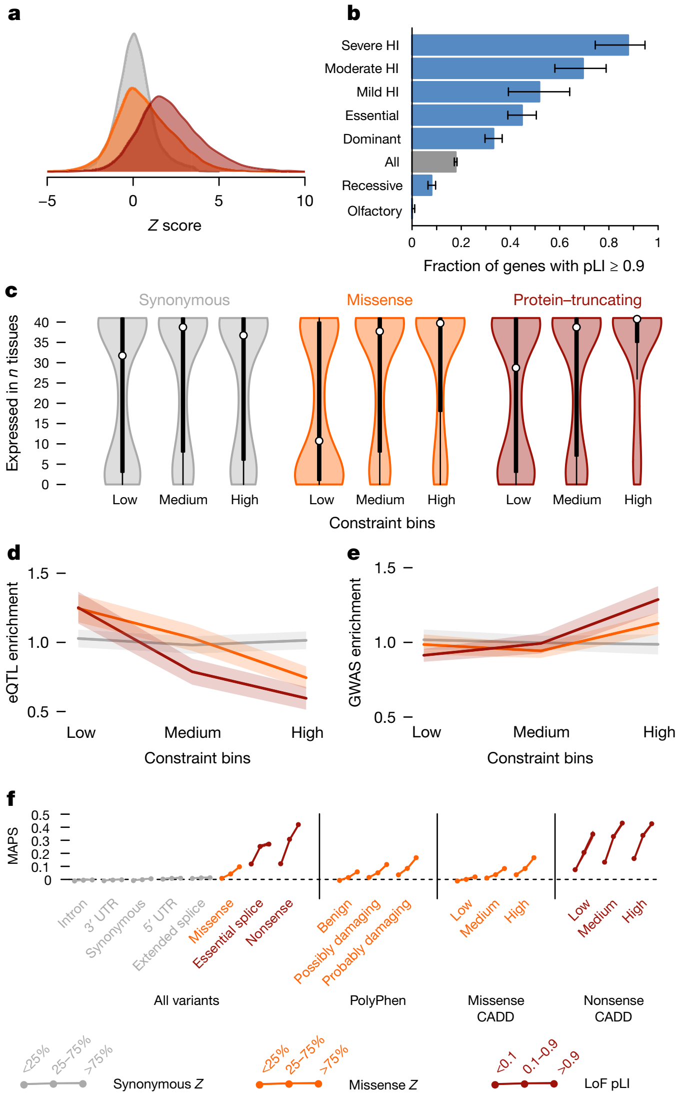


Figure 3 | Quantifying intolerance to functional variation in genes and gene sets. a, Histograms of constraint Z scores for 18,225 genes. This measure of departure of number of variants from expectation is normally distributed for synonymous variants, but right-shifted (higher constraint) for missense and protein-truncating variants (PTVs), indicating that more genes are intolerant to these classes of variation. **b**, The proportion of genes that are very probably intolerant of loss-of-function variation (pLI ≥ 0.9) is highest for ClinGen haploinsufficient (HI) genes, and stratifies by the severity and age of onset of the haploinsufficient phenotype. Genes essential in cell culture and dominant disease genes are likewise enriched for intolerant genes, whereas recessive disease genes and olfactory receptors have fewer intolerant genes. Black error bars indicate 95% confidence intervals. **c**, Synonymous Z scores show no correlation with the number of tissues in which a gene is expressed, but the most missense- and PTV-constrained genes tend to be expressed in more tissues. Thick black bars indicate the first to third quartiles, with the white circle marking the median. **d**, Highly missense- and PTV-constrained genes are less likely to have eQTLs discovered in GTEx as the average gene. Shaded regions around the lines indicate 95% confidence intervals. **e**, Highly missense- and PTV-constrained genes are more likely to be adjacent to genome-wide association study (GWAS) signals than the average gene. Shaded regions around the lines indicate 95% confidence intervals. **f**, MAPS (Fig. 2d) is shown for each functional category, broken down by constraint score bins as shown. Missense and PTV constraint score bins provide information about natural selection at least partially orthogonal to MAPS, PolyPhen, and CADD scores, indicating that this metric should be useful in identifying variants associated with deleterious phenotypes. Shaded regions around the lines indicate 95% confidence intervals. For panels **a**, **c**–**f**, variants are coloured with synonymous in grey, missense in orange, and protein-truncating in maroon.

for missense ($P < 10^{-7}$) and loss-of-function mutations ($P < 10^{-50}$), in line with previous work[15]. The high correlation between the observed and expected number of synonymous variants on the X chromosome ($r = 0.97$ versus 0.98 for autosomes) indicates that this difference in constraint is not due to a calibration issue. To reduce confounding by coding sequence length for PTVs, we developed an expectation-maximization algorithm (Supplementary Information section 4.4) using the observed and expected PTV counts within each gene to separate genes into three categories: null (observed ≈ expected), recessive (observed ≤ 50% of expected), and haploinsufficient (observed < 10% of expected). This metric—the probability of being loss-of-function (LoF) intolerant (pLI)—separates genes of sufficient length into LoF intolerant (pLI ≥ 0.9, $n = 3{,}230$) or LoF tolerant (pLI ≤ 0.1, $n = 10{,}374$) categories. pLI is less correlated with coding sequence length ($r = 0.17$ as compared to 0.57 for the PTV Z score), outperforms the PTV Z score as an intolerance metric (Supplementary Table 15), and reveals the expected contrast between gene lists (Fig. 3b). pLI is positively correlated with the number of physical interaction partners of a gene product ($P < 10^{-41}$). The most constrained pathways (highest median pLI for the genes in the pathway) are core biological processes (spliceosome, ribosome, and proteasome components; Kolmogorov–Smirnov test $P < 10^{-6}$ for all), whereas olfactory receptors are among the least constrained pathways (Kolmogorov–Smirnov test $P < 10^{-16}$), as demonstrated in Fig. 3b, and this is consistent with previous work[5,16–19].

Crucially, we note that LoF-intolerant genes include virtually all known severe haploinsufficient human disease genes (Fig. 3b), but that 72% of LoF-intolerant genes have not yet been assigned a human disease phenotype despite clear evidence for extreme selective constraint (Supplementary Table 13). We note that this extreme constraint does not necessarily reflect a lethal disease or status as a disease gene (for example, *BRCA1* has a pLI of 0), but probably points to genes in which heterozygous loss of function confers some non-trivial survival or reproductive disadvantage.

The most highly constrained missense (top 25% missense Z scores) and PTV (pLI ≥ 0.9) genes show higher expression levels and broader tissue expression than the least constrained genes[20] (Fig. 3c). These most highly constrained genes are also depleted for expression quantitative trait loci (eQTLs) ($P < 10^{-9}$ for missense and PTV; Fig. 3d), yet are enriched within genome-wide significant trait-associated loci (χ^2 test, $P < 10^{-14}$, Fig. 3e). Genes intolerant of PTV variation would be expected to be dosage-sensitive, as in such genes natural selection does not tolerate a 50% deficit in expression due to the loss of single allele. It is thus unsurprising that these genes are also depleted of common genetic variants that have a large enough effect on expression to be detected as eQTLs with current limited sample sizes. However, smaller changes in the expression of these genes, through weaker eQTLs or functional variants, are more likely to contribute to medically relevant phenotypes.

Finally, we investigated how these constraint metrics would stratify mutational classes according to their frequency spectrum, corrected for mutability as in the previous section (Fig. 3f). The effect was most dramatic when considering nonsense variants in the LoF-intolerant set of genes. For missense variants, the missense Z score offers information orthogonal to Polyphen2 and CADD classifications, which are measures predicting the likely deleteriousness of variants, indicating that gene-level measures of constraint offer additional information to variant-level metrics in assessing potential pathogenicity.

ExAC improves variant interpretation in rare disease

We assessed the value of ExAC as a reference data set for clinical sequencing approaches, which typically prioritize or filter potentially deleterious variants on the basis of functional consequence and allele frequency[6]. Filtering on ExAC reduced the number of candidate protein-altering variants by sevenfold compared to the ESP data set, and was most powerful when the highest allele frequency in any one population ('popmax') was used rather than the average

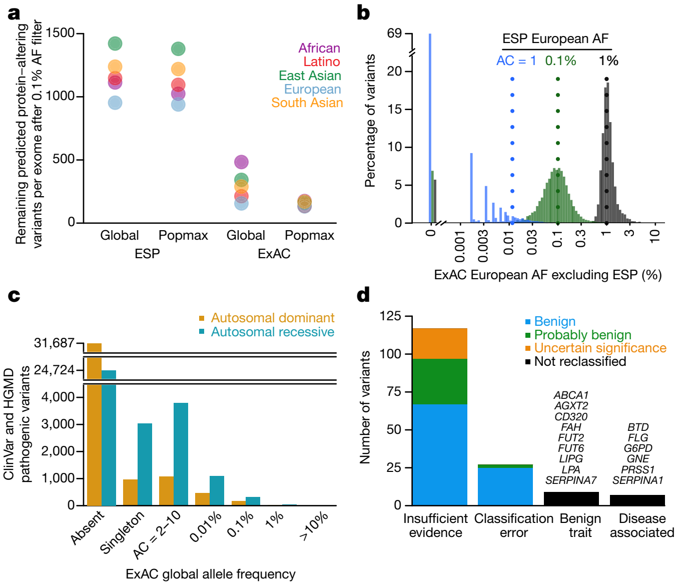


Figure 4 | Filtering for Mendelian variant discovery. a, Predicted missense and protein-truncating variants in 500 randomly chosen ExAC individuals were filtered based on allele frequency (AF) information from ESP, or from the remaining ExAC individuals. At a 0.1% allele frequency filter, ExAC provides greater power to remove candidate variants, leaving an average of 154 variants for analysis, compared to 1,090 after filtering against ESP. Popmax allele frequency also provides greater power than global allele frequency, particularly when populations are unequally sampled. **b**, Estimates of allele frequency in Europeans based on ESP are more precise at higher allele frequencies. Sampling variance and ascertainment bias make allele frequency estimates unreliable, posing problems for Mendelian variant filtration. 69% of ESP European singletons are not seen a second time in ExAC (tall bar at left), illustrating the dangers of filtering on very low allele counts. **c**, Allele frequency spectrum of disease-causing variants in the Human Gene Mutation Database (HGMD) and/or pathogenic or probable pathogenic variants in ClinVar for well-characterized autosomal dominant and autosomal recessive disease genes[28]. Most are not found in ExAC; however, many of the reportedly pathogenic variants found in ExAC are at too high a frequency to be consistent with disease prevalence and penetrance. **d**, Literature review of variants with >1% global allele frequency or >1% Latin American or South Asian population allele frequency confirmed there is insufficient evidence for pathogenicity for the majority of these variants. Variants were reclassified by American College of Medical Genetics and Genomics (ACMG) guidelines[24].

('global') allele frequency (Fig. 4a). ESP is not well-powered to filter at 0.1% allele frequency without removing many genuinely rare variants, as allele frequency estimates based on low allele counts are both upward-biased and imprecise (Fig. 4b). We thus expect that ExAC will provide a very substantial boost in the power and accuracy of variant filtering in Mendelian disease projects.

Previous large-scale sequencing studies have repeatedly shown that some purported Mendelian disease-causing genetic variants are implausibly common in the population[21–23] (Fig. 4c). The average ExAC participant harbours ~54 variants reported as disease-causing in two widely used databases of disease-causing variants (Supplementary Information section 5.2). Most (~41) of these are high-quality genotypes but with implausibly high (>1%) popmax allele frequencies. We therefore hypothesized that most of the supposed burden of Mendelian disease alleles per person is due not to genotyping error, but rather to misclassification in the literature and/or in databases.

We manually curated the evidence of pathogenicity for 192 previously reported pathogenic variants with allele frequency >1% either globally or in South Asian or Latino individuals, populations that are underrepresented in previous reference databases. Nine variants had sufficient data to support disease association, typically with either mild or incompletely penetrant disease effects; the remainder either had insufficient evidence for pathogenicity, no claim of pathogenicity, or were benign traits (Supplementary Information section 5.3). It is difficult to prove the absence of any disease association, and incomplete penetrance or genetic modifiers may contribute in some cases. Nonetheless, the high cumulative allele frequency of these variants combined with their limited original evidence for pathogenicity suggest little contribution to disease, and 163 variants met American College of Medical Genetics criteria[24] for reclassification as benign or probably benign (Fig. 4d). A total of 126 of these 163 have been reclassified in source databases as of December 2015 (Supplementary Table 20). Supporting functional data were reported for 18 of these variants, highlighting the need to review cautiously even variants with experimental support.

We also sought phenotypic data for a subset of ExAC participants homozygous for reported severe recessive disease variants, again enabling reclassification of some variants as benign. North American Indian childhood cirrhosis is a recessive disease of cirrhotic liver failure during childhood requiring liver transplant for survival to adulthood, previously reported to be caused by *CIRH1A* p.R565W[25] (*CIRH1A* is also known as *UTP4*). ExAC ontains 222 heterozygous and 4 homozygous Latino individuals, with a population allele frequency of 1.92%. The 4 homozygotes had no history of liver disease and recontact in two individuals revealed normal liver function (Supplementary Table 22). Thus, despite the rigorous linkage and Sanger sequencing efforts that led to the original report of pathogenicity, the ExAC data demonstrate that this variant is either benign or insufficient to cause disease, highlighting the importance of matched reference populations.

The above curation efforts confirm the importance of allele frequency filtering in analysis of candidate disease variants[6,26,27]. However, literature and database errors are prevalent even at lower allele frequencies: the average ExAC individual contains 0.89 (<1% popmax allele frequency) reportedly Mendelian variants in well-characterized dominant disease genes[28], and 0.21 at <0.1% popmax allele frequency. This inflation probably results from a combination of false reports of pathogenicity and incomplete penetrance, as we have recently shown for *PRNP*[29]. The abundance of rare functional variation in many disease genes in ExAC is a reminder that such variants should not be assumed to be causal or highly penetrant without careful segregation or case-control analysis[7,24].

Effect of rare protein-truncating variants

We investigated the distribution of PTVs, variants predicted to disrupt protein-coding genes through the introduction of a stop codon, frameshift, or the disruption of an essential splice site; such variants are expected to be enriched for complete loss of function of the affected genes. Naturally occurring PTVs in humans provide a model for the functional impact of gene inactivation, and have been used to identify many genes in which LoF causes severe disease[30], as well as rare cases where LoF is protective against disease[31].

Among the 7,404,909 high-quality variants in ExAC, we found 179,774 high-confidence PTVs (as defined in Supplementary Information section 6), 121,309 of which are singletons. This corresponds to an average of 85 heterozygous and 35 homozygous PTVs per individual (Fig. 5a). The diverse nature of the cohort enables the discovery of substantial numbers of new PTVs: out of 58,435 PTVs with an allele count greater than one, 33,625 occur in only one population. However, although PTVs as a category are extremely rare, the majority of the PTVs found in any one person are common, and each individual has only ~2 singleton PTVs, of which 0.14 are found in PTV-constrained genes (pLI > 0.9). ExAC recapitulates known aspects of population demographic models, including an increase in intermediate-frequency (1–5%) PTVs in Finland[32] and relatively common (>1%) PTVs in Africans (Fig. 5b). However, these differences are diminished when considering only LoF-constrained (pLI > 0.9) genes (Extended Data Fig. 4).

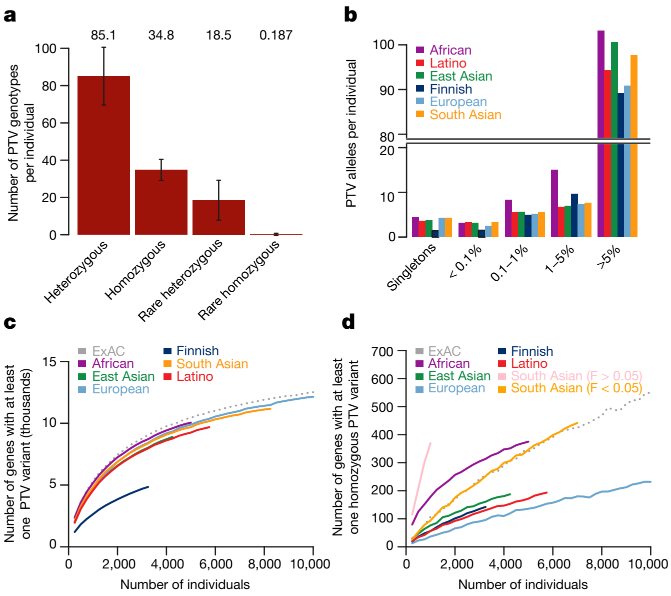


Figure 5 | Protein-truncating variation in ExAC. a, The average ExAC individual has 85 heterozygous and 35 homozygous protein-truncating variants (PTVs), of which 18 and 0.19 are rare (<1% allele frequency), respectively. Error bars represent standard deviation. **b**, Breakdown of PTVs per individual (**a**) by popmax allele frequency bin. Across all populations, most PTVs found in a given individual are common (>5% allele frequency). **c**, **d**, Number of genes with at least one PTV (**c**), or homozygous PTV (**d**), as a function of number of individuals, downsampled from ExAC. The South Asian population is broken down by consanguinity (inbreeding coefficient, *F*). At 60,000 individuals for ExAC, the plots in **c**, **d**, extend to 15,750 with at least one PTV and 1,550 genes with at least one homozygous PTV. Dotted line represents all ExAC samples.

Using a sub-sampling approach, we show that the discovery of both heterozygous (Fig. 5c) and homozygous (Fig. 5d) PTVs scales very differently across human populations, with implications for the design of large-scale sequencing studies to ascertain human knockouts, as described later.

Discussion

Here we describe the generation and analysis of the most comprehensive catalogue (to our knowledge) of human protein-coding genetic variation to date, incorporating high-quality exome sequencing data from 60,706 individuals of diverse geographic ancestry. The resulting call set provides unprecedented resolution for the analysis of low-frequency protein-coding variants in human populations, as well as a public resource (http://exac.broadinstitute.org) for the clinical interpretation of genetic variants observed in disease patients.

The very large sample size of ExAC also provides opportunities for a high-resolution analysis of the sensitivity of human genes to functional variation. Although previous sample sizes have been adequately powered for the assessment of gene-level intolerance to missense variation[11,14], ExAC provides sufficient power for the first time to investigate genic intolerance to PTVs, highlighting 3,230 highly LoF-intolerant genes, 72% of which have no established human disease phenotype in the OMIM or ClinVar databases of observed human genetic mutations. Although this extreme depletion of PTVs will probably highlight genes in which loss of a single copy has been reproductively disadvantageous over recent human history, not all high pLI genes will lead to lethal disease. Additionally, disease genes—particularly those that act after post-reproductive age—do not necessarily have high pLI values (for example, the pLI of *BRCA1* is 0). In separate work[33] we show that ExAC similarly provides power to identify genes intolerant of copy number variation. Quantification of genic intolerance to both classes of variation will provide added power to disease studies.

The ExAC resource provides the largest database to date (to our knowledge) for the estimation of allele frequency for protein-coding genetic variants, providing a powerful filter for analysis of candidate pathogenic variants in severe Mendelian diseases. Frequency data from ESP[1] have been widely used for this purpose, but those data are limited by population diversity and by resolution at allele frequencies ≤ 0.1%. ExAC therefore provides substantially improved power for Mendelian analyses, although it is still limited in power at lower allele frequencies, emphasizing the need for more sophisticated pathogenic variant filtering strategies alongside on-going data aggregation efforts.

We show that different populations confer different advantages in the discovery of gene-disrupting PTVs, providing guidance for the identification of human knockouts to understand gene function. Sampling multiple populations would probably be a fruitful strategy for a researcher investigating common PTV variation. However, discovery of homozygous PTVs is markedly enhanced in the South Asian samples, which come primarily from a Pakistani cohort with 38.3% of individuals self-reporting as having closely related parents, emphasizing the extreme value of consanguineous cohorts for human knockout discovery[34–36] (Fig. 5d). Other approaches to enriching for homozygosity of rare PTVs, such as focusing on bottlenecked populations, have already proved fruitful[32,34].

Even with this large collection of jointly processed exomes, many limitations remain. First, most ExAC individuals were ascertained for biomedically important disease; although we have attempted to exclude severe paediatric diseases, the inclusion of both cases and controls for several polygenic disorders means that ExAC certainly contains disease-associated variants[37]. Second, future reference databases would benefit from including a broader sampling of human diversity, especially from under-represented Middle Eastern and African populations. Third, the inclusion of whole genomes will also be critical to investigate additional classes of functional variation and identify non-coding constrained regions. Finally, and most critically, detailed phenotype data are unavailable for the vast majority of ExAC samples; future initiatives that assemble sequence and clinical data from very large-scale cohorts will be required to fully translate human genetic findings into biological and clinical understanding.

Although the ExAC data set exceeds the scale of previously available frequency reference data sets, much remains to be gained by further increases in sample size. Indeed, the fact that even the rarest transversions have mutational rates[11] on the order of 1×10^{-9} implies that the vast majority of possible non-lethal SNVs probably exist in some living human. ExAC already includes >63% of all possible protein-coding CpG transitions at well-covered synonymous sites; orders-of-magnitude increases in sample size will eventually lead to saturation of other classes of variation.

ExAC was made possible by the willingness of multiple large disease-focused consortia to share their raw data, and by the availability of the software and computational resources required to create a harmonized variant call set on the scale of tens of thousands of samples. The creation of yet larger reference variant databases will require continued emphasis on the value of genomic data sharing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 19 October 2015; accepted 24 June 2016.

1. Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493,** 216–220 (2012).
2. The 1000 Genomes Project Consortium A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
3. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475,** 493–496 (2011).
4. Stoneking, M. & Krause, J. Learning about human population history from ancient and modern genomes. *Nature Rev. Genet.* **12,** 603–614 (2011).

5. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335,** 823–828 (2012).
6. Bamshad, M. J. *et al.* Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Rev. Genet.* **12,** 745–755 (2011).
7. MacArthur, D. G. *et al.* Guidelines for investigating causality of sequence variants in human disease. *Nature* **508,** 469–476 (2014).
8. The Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519,** 223–228 (2015).
9. Fromer, M. *et al. De novo* mutations in schizophrenia implicate synaptic networks. *Nature* **506,** 179–184 (2014).
10. Cooper, D. N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. *Hum. Genet.* **78,** 151–155 (1988).
11. Samocha, K. E. *et al.* A framework for the interpretation of *de novo* mutation in human disease. *Nature Genet.* **46,** 944–950 (2014).
12. Tennessen, J. A. *et al.* Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science* **337,** 64–69 (2012).
13. Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature Genet.* **47,** 435–444 (2015).
14. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet.* **9,** e1003709 (2013).
15. Vicoso, B. & Charlesworth, B. Evolution on the X chromosome: unusual patterns and processes. *Nature Rev. Genet.* **7,** 645–653 (2006).
16. Jeong, H., Mason, S. P., Barabási, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
17. Goh, K.-I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104,** 8685–8690 (2007).
18. Rolland, T. *et al.* A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).
19. Itan, Y. *et al.* The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc. Natl Acad. Sci. USA* **112,** 13615–13620 (2015).
20. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
21. Bell, C. J. *et al.* Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci. Transl. Med.* **3,** 65ra4 (2011).
22. Xue, Y. *et al.* Deleterious- and disease-allele prevalence in healthy individuals: insights from current predictions, mutation databases, and population-scale resequencing. *Am. J. Hum. Genet.* **91,** 1022–1032 (2012).
23. Piton, A., Redin, C. & Mandel, J.-L. XLID-causing mutations and associated genes challenged in light of data from large-scale human exome sequencing. *Am. J. Hum. Genet.* **93,** 368–383 (2013).
24. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17,** 405–423 (2015).
25. Chagnon, P. *et al.* A missense mutation (R565W) in *Cirhin* (FLJ14728) in North American Indian childhood cirrhosis. *Am. J. Hum. Genet.* **71,** 1443–1449 (2002).
26. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133,** 1–9 (2014).
27. Dewey, F. E. *et al.* Sequence to medical phenotypes: a framework for interpretation of human whole genome DNA sequence data. *PLoS Genet.* **11,** e1005496 (2015).
28. Blekhman, R. *et al.* Natural selection on genes that underlie human disease susceptibility. *Curr. Biol.* **18,** 883–889 (2008).
29. Minikel, E. V. *et al.* Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8,** 322ra9 (2016).
30. Chong, J. X. *et al.* The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet.* **97,** 199–215 (2015).
31. Kathiresan, S. Developing medicines that mimic the natural successes of the human genome: lessons from *NPC1L1*, *HMGCR*, *PCSK9*, *APOC3*, and *CETP*. *J. Am. Coll. Cardiol.* **65,** 1562–1566 (2015).
32. Lim, E. T. *et al.* Distribution and medical impact of loss-of-function variants in the Finnish founder population. *PLoS Genet.* **10,** e1004494 (2014).
33. Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nature Genet.* http://dx.doi.org/10.1038/ng.3638 (2016).
34. Sulem, P. *et al.* Identification of a large set of rare complete human knockouts. *Nature Genet.* **47,** 448–452 (2015).
35. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science* http://dx.doi.org/10.1126/science.aac8624 (2016).
36. Saleheen, D. *et al.* Human knockouts in a cohort with a high rate of consanguinity. *Preprint at bioRxiv* http://dx.doi.org/10.1101/031518 (2015).
37. Freischmidt, A. *et al.* Haploinsufficiency of *TBK1* causes familial ALS and fronto-temporal dementia. *Nature Neurosci.* **18,** 631–636 (2015).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank the scientific community for their support and comments on bioRxiv, Twitter, and other public forums, and B. Bulik-Sullivan, J. Bloom and R. Walters for their help with mathematical notation. The full acknowledgements are detailed in Supplementary Information section 8.

Author Contributions M.Le., K.J.K., E.V.M., K.E.S., E.B., T.F., A.H.O., J.S.W., A.J.H., B.B.C., T.T., D.P.B., J.A.K., L.E.D., K.E., F.Z., J.Z., E.P., M.J.D. and D.G.M. contributed to the analysis and writing of the manuscript. M.Le., E.B., T.F., K.J.K., E.V.M., F.Z., D.P.B., J.B., D.N.C., N.D., M.D., R.D., J.F., M.F., L.G., J.G., N.G., D.H., A.K., M.I.K., A.L.M., P.N., L.O., G.M.P., R.P., M.A.R., V.R., S.A.R., D.M.R., K.S., P.D.S., C.S., B.P.T., G.T., M.T.T., B.W., H.W., D.Y., S.B.G., M.J.D. and D.G.M. contributed to the production of the ExAC data set. D.M.A., D.A., M.B., J.D., S.D., R.E., J.C.F., S.B.G., G.G., S.J.G., C.M.H., S.K., M.La., S.M., M.I.M., D.M., R.M., B.M.N., A.P., S.M.P., D.S., J.M.S., P.S., P.F.S., J.T., M.T.T., H.C.W., J.G.W., M.J.D. and D.G.M. contributed to the design and conduct of the various exome sequencing studies and review of the manuscript.

Author Information ExAC data set is publicly available at (http://exac.broadinstitute.org). Reprints and permissions information is available at www.nature.com/reprints. The authors declare competing financial interests: details are available in the online version of the paper. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.G.M. (danmac@broadinstitute.org).

Reviewer Information *Nature* thanks L. Biesecker, J. Shedure and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Variant discovery. We assembled approximately 1 petabyte of raw sequencing data (FASTQ files) from 91,796 individual exomes drawn from a wide range of primarily disease-focused consortia (Supplementary Table 2). We processed these exomes through a single informatic pipeline and performed joint variant calling of single nucleotide variants (SNVs) and indels across all samples using a new version of the Genome Analysis Toolkit (GATK) HaplotypeCaller pipeline. Variant discovery was performed within a defined exome region that includes Gencode v19 coding regions and flanking 50 bases. At each site, sequence information from all individuals was used to assess the evidence for the presence of a variant in each individual. Full details of data processing, variant calling and resources are described in the Supplementary Information sections 1.1–1.4.

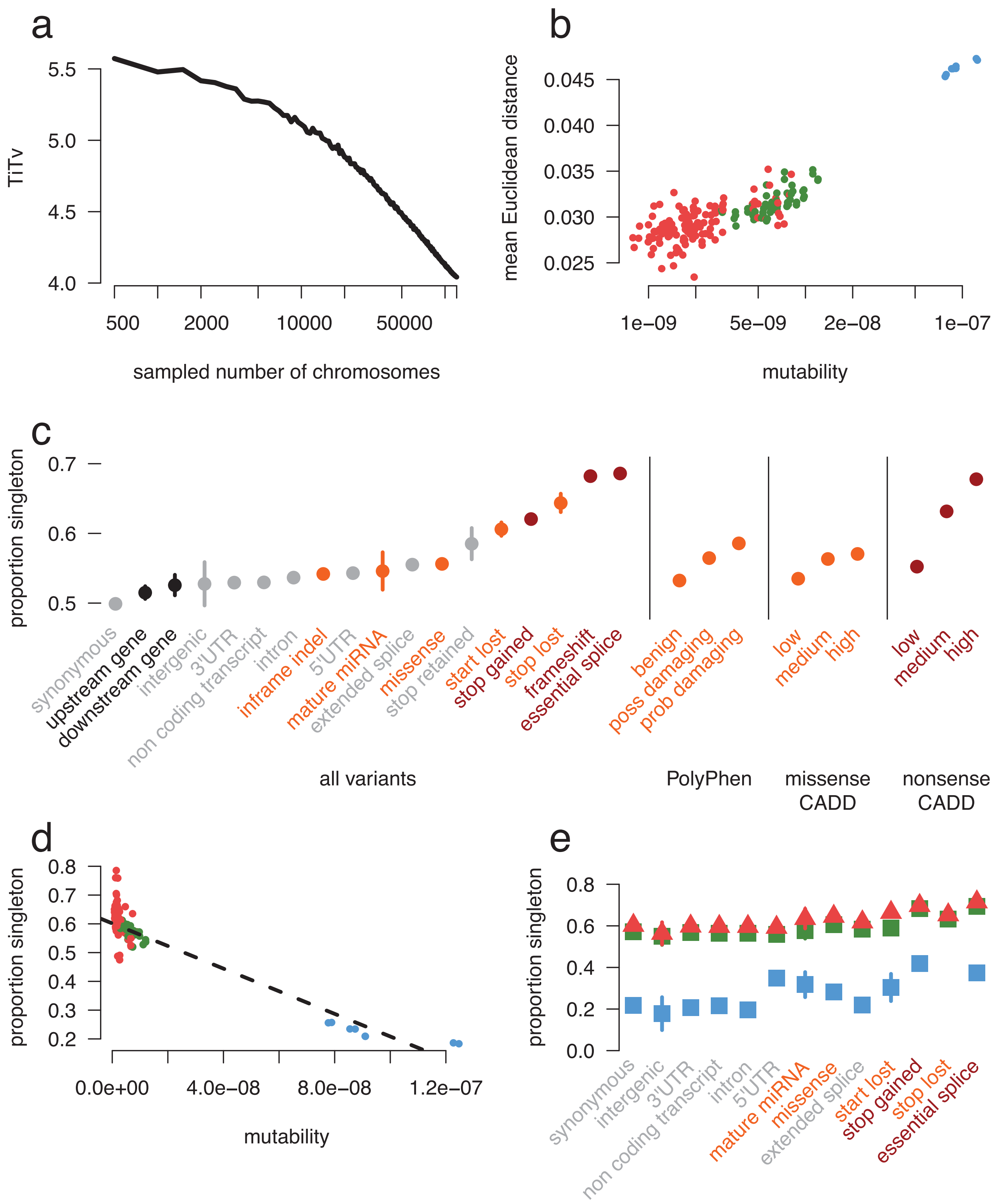
Quality assessment. We leveraged a variety of sources of internal and external validation data to calibrate filters and evaluate the quality of filtered variants (Supplementary Table 7). We adjusted the standard GATK variant site filtering[38] to increase the number of singleton variants that pass this filter, while maintaining a singleton transmission rate of 50.1%, very near the expected 50%, within sequenced trios. We then used the remaining passing variants to assess depth and genotype quality filters compared to >10,000 samples that had been directly genotyped using SNP arrays (Illumina HumanExome) and achieved 97–99% heterozygous concordance, consistent with known error rates for rare variants in chip-based genotyping[39]. Relative to a 'platinum standard' genome sequenced using five different technologies[40], we achieved sensitivity of 99.8% and false discovery rates (FDR) of 0.056% for single nucleotide variants (SNVs), and corresponding rates of 95.1% and 2.17% for insertions and deletions (indels), respectively. Lastly, we compared 13 representative non-Finnish European exomes included in the call set with their corresponding 30× PCR-free genome. The overall SNV and indel FDR was 0.14% and 4.71%, respectively, while for SNV singletons it was 0.389%. The overall FDR by annotation classes missense, synonymous and protein truncating variants (including indels) were 0.076%, 0.055% and 0.471% respectively (Supplementary Tables 5 and 6). Full details of quality assessments are described in Supplementary Information section 1.6.

Sample filtering. The 91,796 samples were filtered based on two criteria. First, samples that were outliers for key metrics were removed (Extended Data Fig. 5b). Second, in order to generate allele frequencies based on independent observations without enrichment of Mendelian disease alleles, we restricted the final release data set to unrelated adults with high-quality sequence data and without severe paediatric disease. After filtering, only 60,706 samples remained, consisting of ~77% of Agilent (33 Mb target) and ~12% of Illumina (37.7 Mb target) exome captures. Full details of the filtering process are described in Supplementary Information section 1.7.

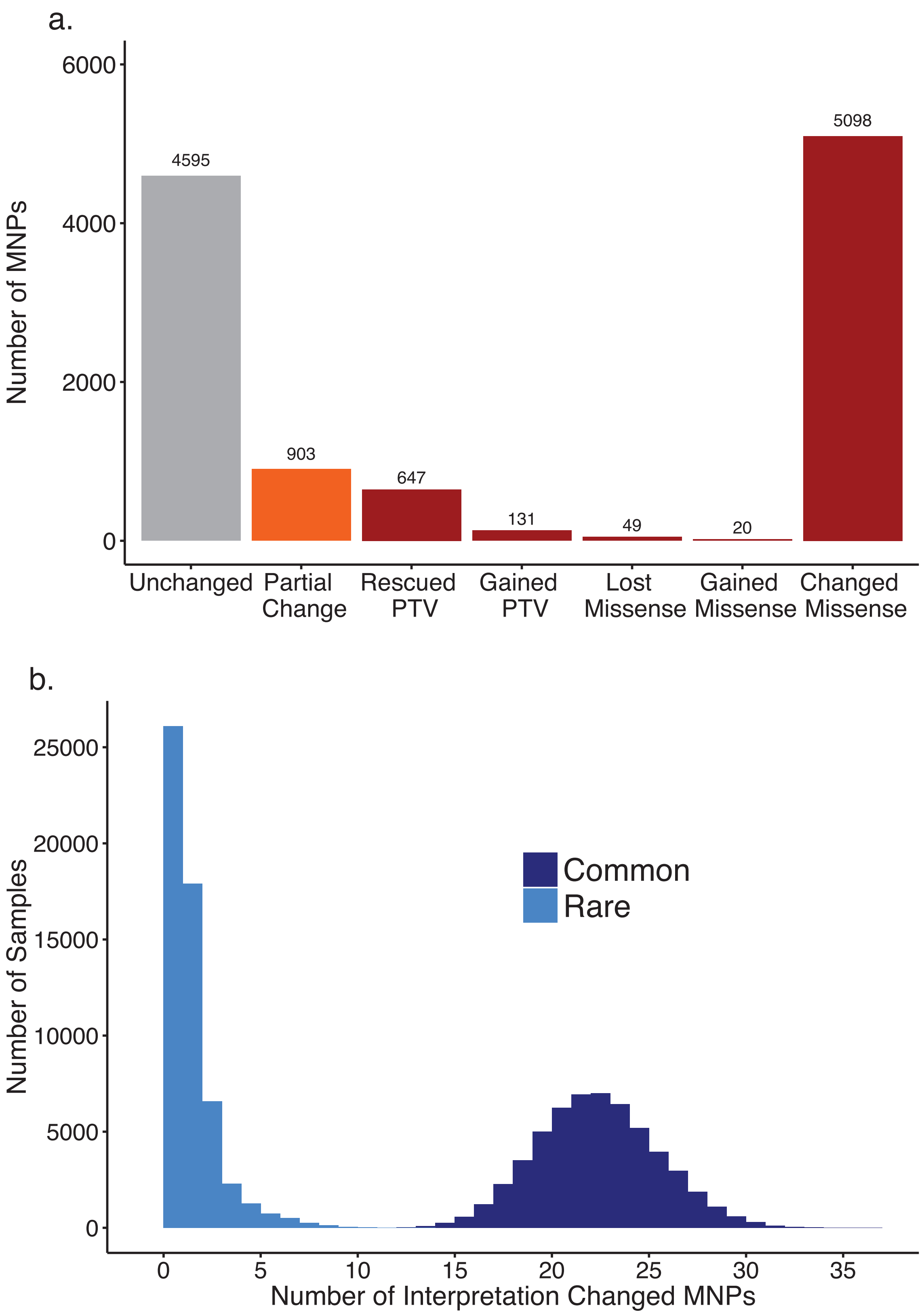
ExAC data release. For each variant, summary data for genotype quality, allele depth and population specific allele counts were calculated before removing all genotype data. This variant summary file was then functionally annotated using variant effect predictor (VEP) with the LOFTEE plugin. This data set can be accessed via the ExAC Browser (http://exac.broadinstitute.org), or downloaded from: (ftp://ftp.broadinstitute.org/pub/ExAC_release/release0.3/ExAC.r0.3.sites.vep.vcf.gz). Full details regarding the annotation of the ExAC data set are described in the Supplementary Information sections 1.9–1.10.

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

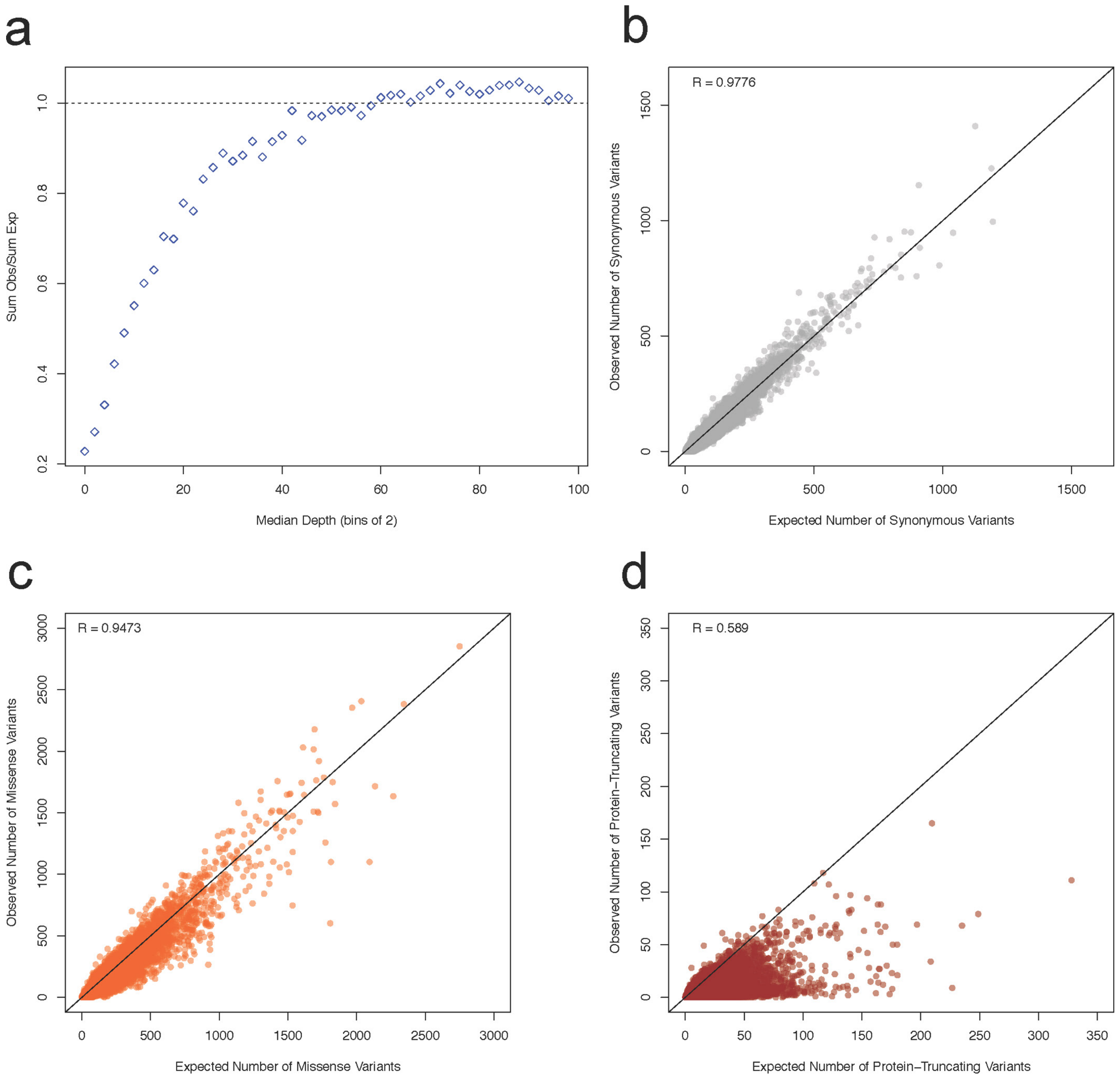
38. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genet.* **43,** 491–498 (2011).
39. Voight, B. F. *et al.* The Metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet.* **8,** e1002793 (2012).
40. Zook, J. M. *et al.* Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nature Biotechnol.* **32,** 246–251 (2014).



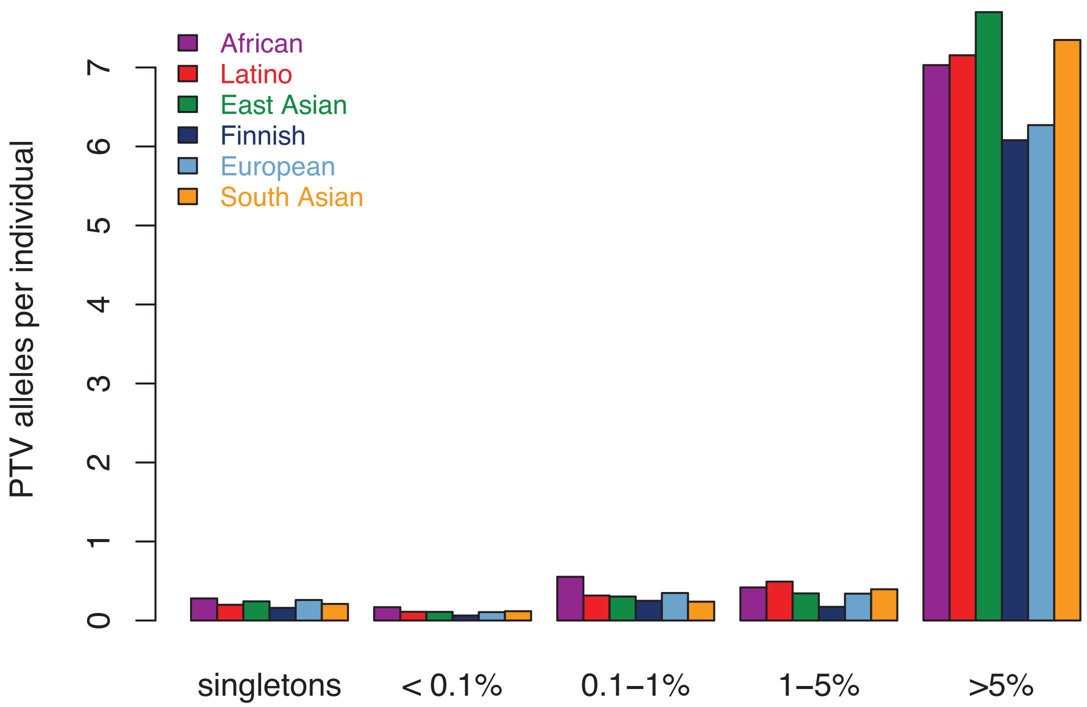
Extended Data Figure 1 | The effect of recurrence across different mutation and functional classes. a, TiTv (transition to transversion) ratio of synonymous variants at downsampled intervals of ExAC. The TiTv is relatively stable at previous sample sizes (<5,000), but changes drastically at larger sample sizes. **b**, For synonymous doubleton variants, mutability of each trinucleotide context is correlated with mean Euclidean distance of individuals that share the doubleton. Transversion (red), and non-CpG transition (green) doubletons are more likely to be found in closer PCA space (more similar ancestries) than CpG transitions (blue). **c**, The proportion singleton among various functional categories. The functional category stop lost has a higher singleton rate than nonsense. Error bars represent standard error of the mean. **d**, Among synonymous variants, mutability of each trinucleotide context is correlated with proportion singleton, suggesting CpG transitions (blue) are more likely to have multiple independent origins driving their allele frequency up. **e**, The proportion singleton metric from **c**, broken down by transversions, non-CpG transitions, and CpG variants. Notably, there is a wide variation in singleton rates among mutational contexts in functional classes, and there are no stop-lost (variants that result in the loss of a stop codon) CpG transitions. Error bars represent standard error of the mean.



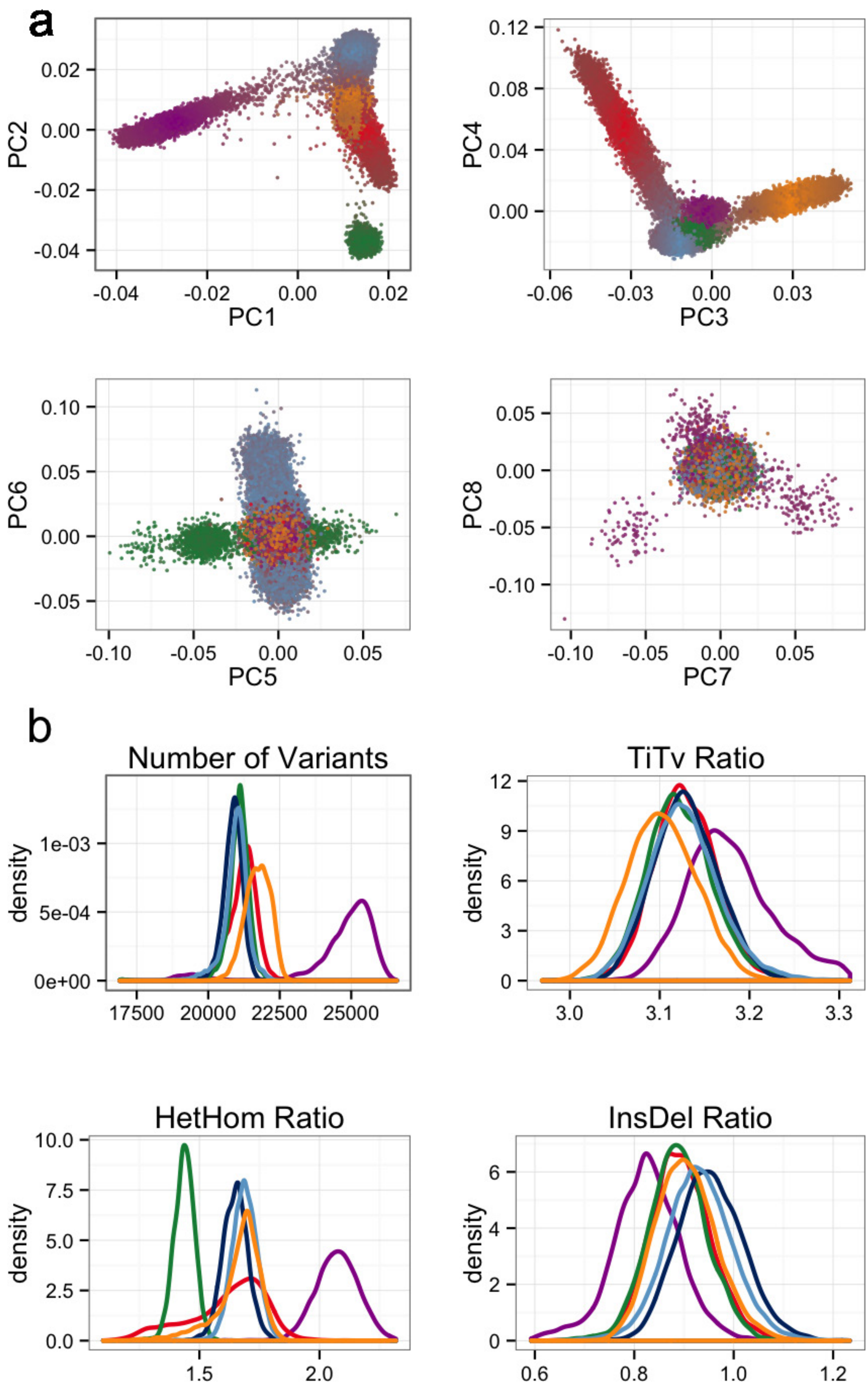
Extended Data Figure 2 | Multi-nucleotide variants discovered in the ExAC data set. a, Number of MNPs per impact on the variant interpretation. **b**, Distribution of the number of MNPs per sample where phasing changes interpretation, separated by allele frequency. Common >1%, rare <1%. MNPs comprised of a rare and common allele are considered rare as this defines the frequency of the MNP.



Extended Data Figure 3 | Relationships between depth and observed versus expected variants, as well as correlations between observed and expected variant counts for synonymous, missense, and protein-truncating. a, The relationship between the median depth of exons (bins of 2) and the sum of all observed synonymous variants in those exons divided by the sum of all expected synonymous variants. The curve was used to determine the appropriate depth adjustment for expected variant counts. For the rest of the panels, the correlation between the depth-adjusted expected variants counts and observed are depicted for synonymous (**b**), missense (**c**), and protein-truncating (**d**). The black line indicates a perfect correlation (slope = 1). Axes have been trimmed to remove *TTN*.



Extended Data Figure 4 | Number of protein-truncating variants in constrained genes per individual by allele frequency bin. Equivalent to Fig. 5b limited to constrained (pLI ≥ 0.9) genes.



Extended Data Figure 5 | Principal component analysis (PCA) and key metrics used to filter samples. a, Principal component analysis using a set of 5,400 common exome SNPs. Individuals are coloured by their distance from each of the population cluster centres using the first 4 principal components. **b**, The metrics number of variants, TiTv, alternate heterozygous/homozygous (HetHom) ratio and indel (InsDel) ratio. Populations are Latino (red), African (purple), European (blue), South Asian (yellow) and East Asian (green).

doi:10.1038/nature19097

Circadian neuron feedback controls the *Drosophila* sleep–activity profile

Fang Guo[1,2], Junwei Yu[2], Hyung Jae Jung[1,2], Katharine C. Abruzzi[1,2], Weifei Luo[1,2], Leslie C. Griffith[2] & Michael Rosbash[1,2]

Little is known about the ability of *Drosophila* circadian neurons to promote sleep. Here we show, using optogenetic manipulation and video recording, that a subset of dorsal clock neurons (DN1s) are potent sleep-promoting cells that release glutamate to directly inhibit key pacemaker neurons. The pacemakers promote morning arousal by activating these DN1s, implying that a late-day feedback circuit drives midday siesta and night-time sleep. To investigate more plastic aspects of the sleep program, we used a calcium assay to monitor and compare the real-time activity of DN1 neurons in freely behaving males and females. Our results revealed that DN1 neurons were more active in males than in females, consistent with the finding that male flies sleep more during the day. DN1 activity is also enhanced by elevated temperature, consistent with the ability of higher temperatures to increase sleep. These new approaches indicate that DN1s have a major effect on the fly sleep–wake profile and integrate environmental information with the circadian molecular program.

Mammalian circadian feedback loops take place in many tissues. They include the suprachiasmatic nucleus (SCN), the ~10,000 neurons of the master pacemaker in the hypothalamus[1,2]. The equivalent circadian region of *Drosophila* brain contains about 75 pairs of neurons; they are arranged in several groups[3] and have a major role in determining the characteristic locomotor activity program[4–6]. It is characterized by morning (M) and evening (E) activity peaks under 12:12 light:dark conditions. There is also a midday siesta between the two activity peaks as well as consolidated sleep at night[7]. M activity is largely determined by the four circadian M cells, the small ventrolateral neurons (sLN_vs)[7,8], whereas E activity is due to three CRY-positive dorsal lateral neurons and the 5th sLN_v (E cells)[9–12]. Although fly sleep is regulated by the clock, there are no known circadian neurons that function predominantly to inhibit locomotor activity or promote sleep, that is, that make a major contribution to the midday siesta or night-time sleep.

In the course of applying different GAL4 lines, optogenetics and a new calcium assay to the study of fly behaviour and circadian neuronal activity, we discovered that a group of glutamatergic dorsal clock neurons (DN1s) are sleep-promoting. Previous work had shown that DN1s function as activity-promoting neurons[5,13], but our results indicate an additional role: glutamatergic DN1s negatively feedback onto M and E cells and thereby promote sleep, especially during midday. Without these neurons and this feedback mechanism, the classical activity/sleep pattern of *Drosophila* is compromised. Our methods also show that these same clock neurons shape the sleep pattern in response to environmental change and should be widely applicable to other fly neurons and behaviours.

DN1 activity shapes the activity and sleep profiles

To monitor more precisely the movement and sleep of adult flies, we used an automated video recording assay instead of DAM (*Drosophila* activity monitor, Trikinetics)[14,15]. We also introduced the use of 96-well plates to allow other experimental manipulations (see Methods and Extended Data Fig. 1; also see ref. 16). In this format, the flies had normal bimodal locomotor activity and stable sleep–wake cycles over many light:dark days (Extended Data Fig. 1). To validate the system, we compared video recording between 96-well plates and Trikinetics tubes; the two methods produced identical patterns (for example, compare Fig. 1a right with centre).

To address the function of DN1s, we first expressed the synaptic neurotransmitter blocker tetanus toxin light chain (TNT) in male flies with the two most commonly used DN1 drivers (*Clk4.1M-GAL4* and *R18H11-GAL4*)[13,17–20]. Comparing these two expression patterns confirmed that *R18H11* promoter is expressed more strongly in a subgroup defined by expression of *Clk4.1M* (Extended Data Fig. 2; for simplicity we will refer to these cells as DN1s). Expression of the inactive toxin (Tet) did not alter wild-type behavioural profiles (data not shown).

The *DN1>TNT* flies (TNT expressed in the DN1s) were more active than control flies at almost all times of day, which markedly reduced the bimodal activity pattern (Fig. 1a). Blocking DN1 neurotransmitter release also strongly decreased the siesta and night-time sleep levels (Fig. 1a and Extended Data Fig. 3a, b). Interestingly, the decrease in total sleep levels of *DN1>TNT* flies was due to a reduction of sleep-episode duration during both the day and night; there was also a slight decrease in locomotor activity during wake (Extended Data Fig. 3a–f). As these DN1-blocked flies were still rhythmic in DD (constant darkness) (Extended Data Fig. 3g), free-running rhythmicity does not require DN1 neurotransmitter output[18]. The data taken together suggest that a DN1 neurotransmitter shapes the standard *Drosophila* light–dark locomotor activity pattern and also enhances sleep levels, a surprising result given the previous role of DN1s in enhancing morning arousal[5,13,21,22].

If blocking DN1 output suppresses sleep, DN1 activation by the red-shifted channelrhodopsin CsChrimson[23,24] should promote sleep and inhibit locomotor activity. To address this possibility, we combined optogenetic stimulation[25] with behavioural monitoring in the 96-well plate format. The light-emitting diode (LED) stimulation ($0.08\ \text{mW mm}^{-2}$) was turned on between zeitgeber time (ZT) 7–24 (ZT is time in LD 12:12), to examine the effect of DN1 activation on siesta, evening peak and night-time sleep.

Red-light-mediated DN1 activation strongly and rapidly affected fly behaviour (Fig. 1b): locomotor activity was suppressed (Fig. 1b, top), and the siesta was extended (Fig. 1c, top). In contrast, the locomotor

[1] Howard Hughes Medical Institute and National Center for Behavioral Genomics, Brandeis University, Waltham, Massachusetts 02454, USA. [2] Department of Biology, National Center for Behavioral Genomics and Volen National Center for Complex Systems, Brandeis University, Waltham, Massachusetts 02454, USA.

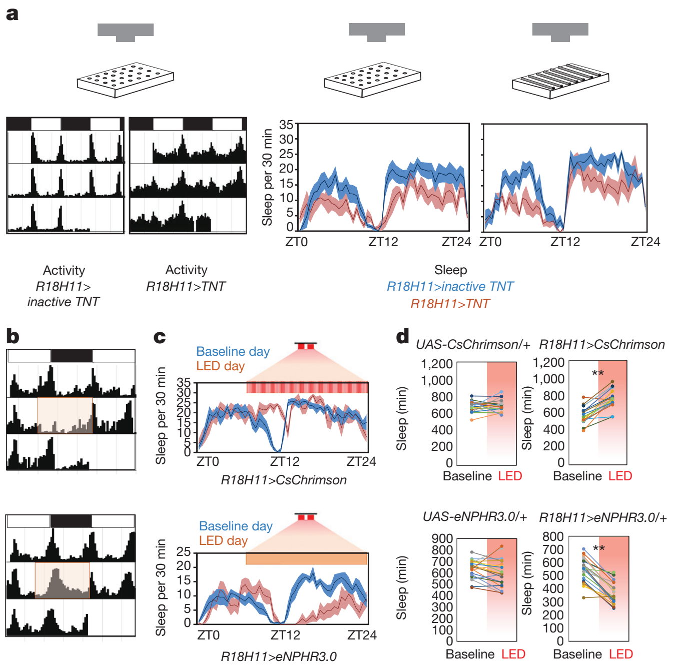


Figure 1 | Manipulation of DN1 activity affects the activity–sleep pattern. a, Activity and sleep data for experimental (*R18H11-GAL4/UAS-TNT*) and control (*R18H11-GAL4/UAS*-inactive *TNT*) male flies. The recording paradigm used (video of 96-well plates or DAM tubes) is indicated by the cartoon. Left, activity data for control (left) and experimental (right) male flies in LD. White bars indicate daytime (ZT 0–12); black bars indicate night-time (ZT 12–24). Right, sleep traces for control (blue) and experimental groups (red). The genotypes are indicated below the panel. Shading corresponds to s.e.m. $n = 14$ *GAL4>inactive TNT* and 16 *GAL4>TNT*. **b**–**d**, Stimulation and inhibition of DN1 firing modulate the E activity peak and sleep. **b**, Activity record of experimental flies. The pink boxes denote the red light (637 nm) stimulation window. **c**, Sleep traces from the baseline day (blue) and the red LED light stimulation day (red) of experimental flies. The pink bar represents the red light illumination (stripe bar upper, 10 HZ light pulse; solid bar lower, constant light). Shading corresponds to s.e.m. **d**, Quantification of sleep gained and lost between the baseline day (white background) and LED stimulation day (pink background) during ZT 7–12 (LED on times) in each group. $n = 19$ *UAS-CsChrimson/+* and *UAS-eNPHR3.0/+*, $n = 20$ *R18H11>CsChrimson* and 21 *R18H11>eNPHR3.0*. Results with shading are mean ± s.e.m. $^{**}P < 0.001$ by paired *t*-test.

activity and sleep levels of flies without CsChrimson expression were similar to the preceding baseline days (Fig. 1d, top left). Addition of *CRY-GAL80* to the *R18H11-GAL4* inhibits DN1 expression (Extended Data Fig. 4a), and expression of CsChrimson with this *R18H11-GAL4;CRY-GAL80* driver did not promote sleep even under much longer 24 h LED stimulation (Extended Data Fig. 4b). In addition, co-expression of TNT and CsChrimson in DN1s reduces sleep and eliminates the sleep-promoting effect of activation, further indicating that DN1s are a source of a sleep-promoting neurotransmitter (Extended Data Fig. 4b).

Importantly, DN1-activated flies have an enhanced arousal threshold as expected from the increased sleep. We assayed arousal threshold by video recording the trajectory of flies in response to a mechanical stimulus (Extended Data Fig. 5). Without LED illumination, a low stimulus (1 tap) woke up ~50% of *DN1>CsChrimson* flies at ZT 6 and 100% of these flies at ZT10. This indicates that they have a higher arousal threshold during siesta than during evening activity (Extended Data Fig. 5b). Significantly, red-light-mediated DN1 activation strongly reduced the percentage of *DN1>CsChrimson* flies aroused by the stimulus at both time points (Extended Data Fig. 5b). The enhanced quiescence was not due to a loss of locomotor ability or a comatose-like state, as a stronger stimulus (10 taps) increased the percentage of aroused flies from ~20% to ~70% (Extended Data Fig. 5b). Furthermore, *DN1>TNT* flies not only had a reduced arousal threshold but were also more sensitive to the low stimulus (1 tap) during siesta time (Extended Data Fig. 5b).

We extended the optogenetic approach by activating the inhibitory halorhodopsin eNpHR3.0. Because eNpHR3.0 responds to constant 630 nm red light[26], we exposed flies to constant high intensity 627 nm light (1 mW mm^{-2}) between ZT 7–24. The activity of control flies is affected by the strong illumination, with a delayed E peak but without a net effect on total sleep; this is because the illumination increased daytime sleep but reduced night-time sleep (data not shown). In contrast, illuminating the *DN1>eNpHR3.0* flies extended the E activity peak throughout the whole night (Fig. 1b, bottom left) and strongly reduced sleep (Fig. 1c–d, bottom right). When the LED was turned on only during the daytime, the siesta was enhanced in control flies as expected from the above results[27], but the *DN1>eNpHR3.0* flies had significantly reduced siesta (Extended Data Fig. 6b). The optogenetic strategies taken together show that the DN1s have a major role in shaping the standard *Drosophila* light–dark locomotor activity pattern and in determining proper sleep levels.

DN1s mediate negative feedback onto core pacemakers

Given the important role of the DN1s on the E peak and sleep, we considered that the DN1s might interact with the activity-promoting core pacemakers, that is, the M and E cells[10]. Because optogenetically activating E cells not only causes immediate activity but also inhibits sleep (data not shown), this opposite behavioural response from activating DN1s suggests an inhibitory interaction between DN1s and these circadian neurons. As the dendritic region of the E cells and the presynaptic region of DN1s are localized to the same brain region, the interaction between these two groups might be direct (Extended Data Fig. 7).

Indeed, GRASP-labelling (GFP reconstitution across synaptic partners[28]) verifies that DN1s and E cells are in close proximity (Fig. 2a green, upper and middle panels). This contact occurs within the E-cell dendritic region as expected (Extended Data Fig. 7). The same GRASP strategy shows that DN1s also contact the dorsal axon region of PDF cells (Fig. 2a) as shown[5,22]. On the basis of previous indications that E cells promote total activity levels[10], that the M cells promote morning activity[18,19] and that the DN1s inhibit locomotor activity and promote sleep, these contacts may reflect inhibitory interactions between DN1s and both M and E cells

To address this possibility, we expressed the adenosine triphosphate (ATP)-gated cation channel P2X2 in DN1s and used GCaMP6f to detect calcium changes in M and E cells after ATP addition to an *in vitro* brain preparations[29,30]. To examine more easily any inhibitory effect of DN1 activation, we performed these experiments at dawn and dusk, when M and E cells have higher neuronal activity and brighter endogenous GCaMP6f signal (unpublished data). ATP perfusion triggers a calcium increase in the DN1 region as expected from P2X2 channel expression (Fig. 2b top and Supplementary Video 1), and there was a simultaneous reduction of GCaMP6f signal in the cell bodies of M and E cells as well as in the dorsal terminal of M cells (Fig. 2b and Extended Data Fig. 8a, b). The results confirm that DN1s have an inhibitory effect on activity-promoting circadian neurons and suggest that this effect transitions flies from 'wake' to 'sleep' under circadian control.

Glutamate modulates the E peak

We next investigated how these DN1 inhibitory interactions modulate the locomotor activity and sleep profiles. Immunostaining indicates that the DN1 projections and especially those of the *R18H11-GAL4*-labelled subset are strongly stained by VGLUT (the vesicular glutamate transporter) antibodies[31] (Supplementary Video 2). Co-staining of *R18H11-LexA* and *VGlut^{MI04979}-GAL4*-labelled neurons confirmed that all of these DN1s are glutamatergic[32] (Fig. 3a). Previous studies

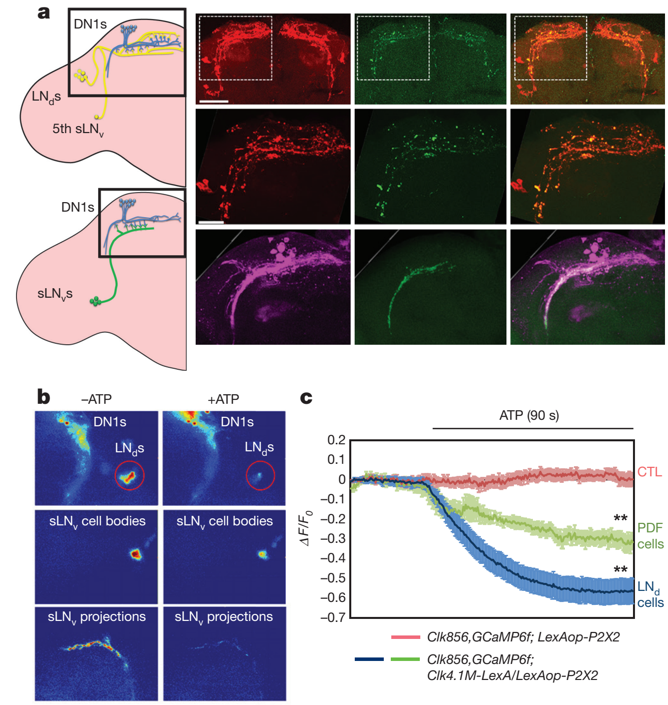


Figure 2 | DN1s directly contact and reduce calcium levels in core pacemakers. a, GRASP assays indicate contact between E cells, M cells and DN1s in the dorsal brain. Red signal indicates the large fragment of GFP (GFP1–10) expressed in the projections of E cells. The green signal shows the contact area between E cells and DN1s. The overlay is shown in the rightmost panel. Scale bar, 50 μm. Magnified images of the boxed area are shown in the middle. Scale bar, 20 μm. The GRASP signal (green) between DN1s (magenta) and M cells is shown in the lower row. **b**, **c**, DN1s inhibit calcium levels in the core pacemakers. **b**, Pseudo-coloured images of GCaMP6f fluorescence intensity from representative brains before and after ATP application. **c**, Mean GCaMP6f response traces are plotted. The genotypes for each group are labelled below. Solid line, ATP application. $n = 8$ for negative control, 7 for PDF cells and 11 for dorsal lateral neurons. Results in **c** are mean ± s.e.m. $^{**}P < 0.001$ by unpaired two-tailed Student's t-test.

have shown that glutamate often functions as an inhibitory neurotransmitter in the fly central nervous system[31,33] (Extended Data Fig. 8c, d). Moreover, RNA profiling of DN1s[34] shows that they are enriched by more than two orders of magnitude in *vglut* mRNA compared to other circadian neuron subgroups (data not shown).

Consistent with these indications, GCaMP6- and Arclight-mediated imaging shows that direct perfusion of glutamate significantly decreased calcium levels and hyperpolarized the membrane potential of both M and E cells (Fig. 3c and Extended Data Fig. 8c, d; E cell data are not shown). Moreover, flies co-expressing *dTrpA1* and *VGlut* RNA interference (RNAi) constructs within the DN1s maintain a higher E activity peak at high temperature than *DN1>dTrpA1* flies (Extended Data Fig. 9), further indicating that glutamate is a DN1-derived inhibitory neurotransmitter (Fig. 2b).

RNA profiling of E cells indicates that they express the metabotropic glutamate receptor mGluRA, which decreases intracellular calcium[31]. Moreover, purification of E cells at different circadian times[34] indicates that this mRNA undergoes strong circadian oscillations with a peak around ZT 7–14 and a trough during the night (Fig. 3b). Cycling of this mRNA can explain why activation of DN1s inhibits E-cell-derived locomotor activity predominantly in the late daytime to early night and is consistent with suggestions from previous studies[31,35,36]. Other DN1-derived neurotransmitters or neuropeptides may be dominant at other times of day, for example, to promote morning activity at dawn[13] (see Discussion).

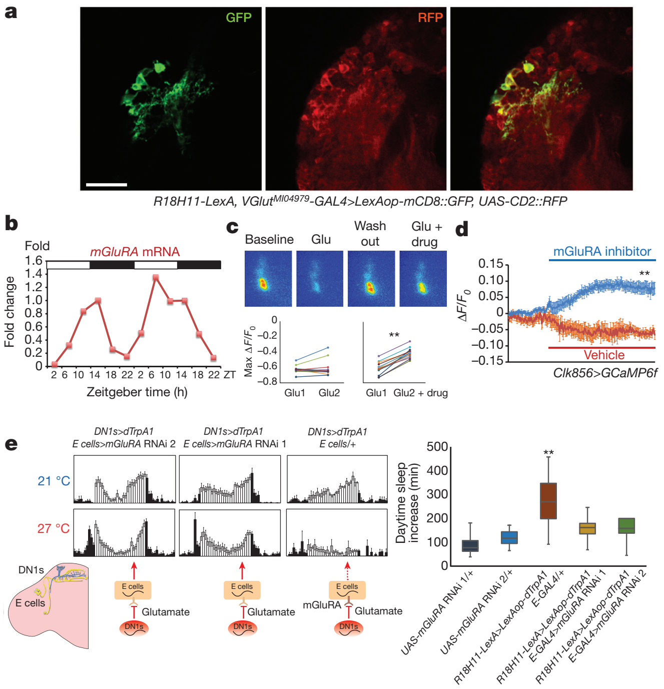


Figure 3 | DN1s inhibit E cells via glutamate release to modulate the E peak and sleep. a, *R18H11*-driver labelled DN1s are glutamatergic. Anti-GFP (left) and anti-RFP (middle) staining was visualized in the dorsal brain of flies. Genotype is shown below. Scale bar, 20 μm. **b**, *mGluRA* mRNA levels cycle in E cells, peaking at midday. Two independent sets of 6 time points for *mGluRA* levels are plotted consecutively. **c**, Co-application of the mGluRA-specific antagonist LY 341495 with glutamate blocks the glutamate-induced inhibition. The 4 middle panels are representative GCaMP6f images of sLN$_v$s during baseline, glutamate perfusion, wash out and glutamate plus LY 341495 perfusion (from left to right). The quantification of peak GCaMP6f changes is plotted in the two lower panels. $n = 8$ for left panel and $n = 11$ for right. $^{**}P < 0.001$ by paired t-test. **d**, Perfusion of LY 341495 increases calcium levels within sLN$_v$s. Mean GCaMP6f response traces are plotted. $n = 11$ for the experimental group and $n = 7$ for the negative control group. $^{**}P < 0.001$ by unpaired two-tailed Student's t-test. Results in panels are mean ± s.e.m. **e**, Reducing mGluRA within E cells impairs the DN1 activation effect on E peak and sleep. Left, the activity pattern of control and *mGluRA* RNAi flies at 21 °C and 27 °C. Right, quantification of daytime sleep changes from different groups. Box boundaries represent the first and third quartiles, and whiskers are 1.5 interquartile range. Genotypes are shown below. $n = 31$, 32, 32, 30 and 29 for genotypes in **e**, respectively. $^{**}P < 0.001$ by one-way ANOVA with Tukey's post-hoc test.

To test further the role of mGluRA, we directly applied the mGluRA-specific inhibitor LY 341495 (700 nM) to dissected fly brains expressing GCaMP6f in circadian neurons[31,33]. It significantly reduced the glutamate-induced calcium decrease in core pacemakers (Fig. 3c). Baseline calcium levels of sLN$_v$s were also modestly but significantly increased (10%) by perfusing LY 341495 (Fig. 3d), suggesting that core pacemaker is inhibited by endogenous glutamate.

An RNAi strategy was used to address the importance of mGluRA expression within circadian cells. Expression of *mGluRA* RNAi in M and E cells with *DvPdf-GAL4* not only decreased the inhibitory effect of glutamate (Extended Data Fig. 10a) but also significantly reduced baseline sleep levels, especially the siesta (Extended Data Fig. 10b). In addition, DN1 activation by dTrpA1 increased daytime sleep and inhibited the E peak, an effect that was blunted by reducing mGluRA levels in E cells (Fig. 3e). Although the RNAi results do not prove that the mRNA cycling is significant, they support a functional glutamate-mediated inhibitory connection between DN1s and E cells. These data further indicate that the DN1s have a major influence on the locomotor activity pattern, by promoting the siesta in a temporally gated manner.

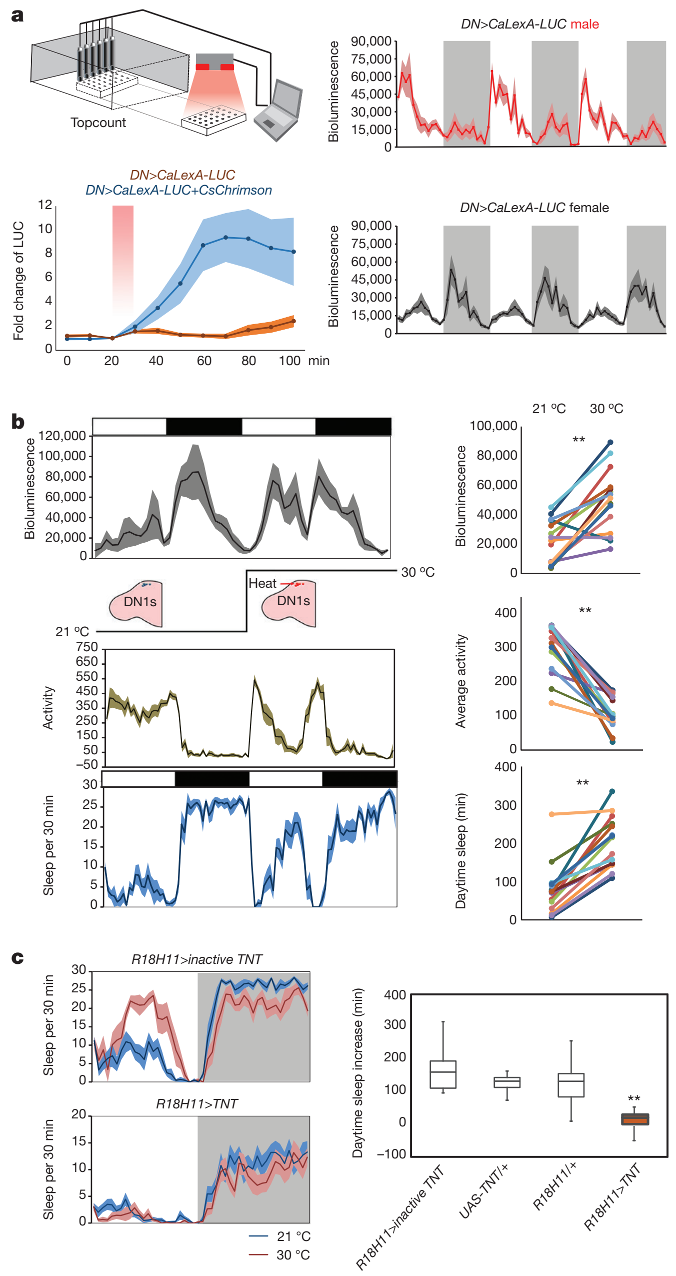


Figure 4 | DN1 neuronal activity is sexually dimorphic and can be activated by warm temperatures to enhance fly sleep. a, Characterization of CaLexA–LUC in freely behaving animals. Left, LUC levels reflect neuronal activity in DN1s after CsChrimson stimulation (lower panel). The fold-change of luminescence was calculated as the ratio of the luminescence level after CsChrimson activation to the baseline luminescence level. The red-shaded box indicates the 10 min 627 nm light pulse. The genotypes of each line are shown below and $n=16$ for each groups. Results with shading represents mean ± s.e.m. Right, CaLexA–LUC shows a marked male–female difference in DN1 activity. Averaged bioluminescence levels of 24 *CLK4.1M-GAL4>CaLexA-LUC* males (red) and females (grey) are plotted. Shaded background depicts dark periods. **b**, The real-time CaLexA–LUC assay reveals that warmer temperatures promote DN1 activity in the daytime. Black boxes indicate dark periods, white boxes indicate light periods. Flies were maintained at 21 °C and then transferred to 30 °C. Shading corresponds to s.e.m. Bioluminescence (arbitrary units); locomotor activity and daytime sleep profile are plotted (left) and quantified (right). $n=15$ for each group; $^{**}P<0.001$ by paired *t*-test. **c**, Blocking DN1 output abolished warm-temperature-induced siesta in females. Sleep traces of control and experimental flies are shown on the left. Blue colour indicates data at 21 °C, and red colour indicates data at 30 °C. The box plot on right shows the sleep increase for the different groups. Box boundaries represent the first and third quartiles, whiskers are 1.5 interquartile range. $n=32$ for each group and shading represent s.e.m. The genotype for each groups are labelled bellow. ** $P<0.001$ by Kruskall–Wallis non-parametric one-way ANOVA with Dunn's multiple comparisons test.

Temperature and sex regulate DN1 activity and sleep

Females have a markedly different locomotor activity and sleep pattern than males: females manifest a much less robust siesta and a less pronounced evening peak. The siesta and evening peak phenotypes are a result of more uniform female daytime activity[37] (Extended Data Fig. 1a). In addition, higher temperatures increase the magnitude of the siesta, which is most apparent in females because of their reduced siesta[38,39]. The robust temperature-dependent increase in the siesta may be an adaptation to seasonal changes, that is, more summer-like conditions[40].

To address this issue in detail, we developed a real-time neuronal activity assay of live flies in the 96-well format. A recently generated calcium-dependent transcription activator *UAS-CaLexA*[41] and *LexAop-LUC* (luciferase) were expressed in DN1s, and individual flies assayed in a standard Topcount plate-reader[42]. LUC activity in these animals should reflect neuronal activity (via calcium levels) in the cells expressing CaLexA (calcium-dependent nuclear import of LexA). We used optogenetics as an initial test of this approach, by co-expressing the CsChrimson with CaLexA–LUC in DN1s and exposing the flies to a 10-min red light pulse; it caused a rapid and marked increase in LUC activity (Fig. 4a). We also tested the inhibitory effect of eNpHR3.0 expression in DN1s, by measuring CaLexA–LUC levels under constant strong red light illumination. The intense light significantly reduced LUC activity for hours (Extended Data Fig. 6a).

Consistent with the difference in siesta between males and females, there is an equally notable difference in the pattern and amplitude of DN1 activity between males and females, as assayed with CaLexA–LUC (Fig. 4a). Male activity increases before light on, peaks during the morning and then declines to a trough in the evening. These patterns and their sexual dimorphism are DN1 specific: CaLexA–LUC male and female patterns from other neurons are completely different, for example, from neurons associated with the ellipsoid body or the ventral fan-shaped body (data not shown). The much higher morning and midday DN1 activity of males probably contributes to their more robust morning anticipation and siesta[13,18,19].

To further probe the relationship of DN1 activity to the siesta, we assayed the response of female DN1 activity to temperature with CaLexA–LUC. There is a marked increase in luciferase activity in the middle of the day at 30 °C compared to 21 °C (Fig. 4b top), which coincides with a prominent temperature-mediated increase in the female siesta and decrease in daytime activity (Fig. 4b middle and bottom). CaLexA–LUC expression in other non-circadian neurons displayed reduced overall activity with the same temperature increase (data not shown), indicating that the DN1 temperature response is specific. To show that DN1 output is necessary for the temperature-dependent siesta increase, we expressed TNT in DN1s and observed little daily sleep increase at 30 °C (Fig. 4c). We therefore suggest that temperature enhances DN1 firing, which promotes the siesta. The data taken together indicate that the DN1s promote the siesta and sleep more generally in a clock-, temperature- and sex-dependent manner.

Conclusions

Although DN1s are activated by M cells[5,22], our data indicate that there is inhibitory feedback by the DN1s onto M and E cells later in the day to promote the siesta and night-time sleep (Extended Data Fig. 10c). This feedback at the level of neuronal circuitry parallels and even exploits the transcriptional negative feedback loop that governs intracellular circadian rhythmicity, to time the siesta and maintain a robust sleep–activity pattern[43,44].

Another group activated the same *R18H11-GAL4*-labelled DN1s and emphasized the wake-promoting effects of DN1 before dawn[13]. However, a positive effect on the midday siesta as well as inhibition of the subsequent E peak as reported here is also evident in their data (not commented on, but in figure 4 of ref. 13). As there is strong evidence that the DN1 firing rate is maximal around the morning in light:dark[13,21], the strong and unanticipated sleep effects of TNT expression (Fig. 1a) and of channelrhodopsin activation (Fig. 1b) are discordant with the morning peak of these DN1 firing rates. (There is no evidence of heterogeneity in the male DN1 electrophysiological data[21], and the CaLexA–LUC activity patterns are weaker but qualitatively similar with the *R18H11-GAL4* driver (data not shown).) This discrepancy implicates cycling signalling molecules (for example, Fig. 3b) in this expansion of the DN1 behavioural repertoire. Moreover, we speculate that the cycling of these same signalling molecules contributes to the delay required for an effective negative feedback circuit (Extended Data Fig. 10c).

The CaLexA assay indicated that DN1 neuronal activity is sexually dimorphic (see below) as well as temperature sensitive. This second feature parallels the known positive effect of temperature on daytime sleep and may reflect a temperature sensor within DN1s or elsewhere within the brain; the adjacent DN2s are good candidates[45]. The siesta increase with temperature may also be related to temperature-sensitive splicing of the *period* (*per*) gene[38,39]. *per* splicing may even occur within DN1s and cause enhanced neuronal activity.

Video monitoring of behaviour in a 96-well format is standard in the zebrafish community[46], and its value for fly-sleep monitoring has been noted elsewhere[14,15]. Although our video results are generally very similar to the activity and sleep profiles from DAM boards, small movements away from the DAM board infrared beam are only detected by video recording and affect the night-time sleep results in the TNT experiments (Fig. 1a).

The CaLexA–LUC assay may be superior for many purposes to recording neuronal activity using other methods, for example electrophysiology or calcium imaging. Indeed, our results indicate substantial differences from dissected brains as well as from calcium imaging of tethered flies[21,47], suggesting that the wake-behaving format is relevant to circadian neuron firing patterns.

This assay also indicates a marked difference between male and female DN1s. To our knowledge, this is the first indication of sexual dimorphism in the fly circadian system. Although we do not know how the CaLexA–LUC signal translates into precise calcium levels or firing rates, the low daytime activity of female DN1s is probably relevant to their relatively weak siesta and morning activity; their high night-time activity suggests a contribution to female-specific night-time tasks. We suggest that this CaLexA–LUC assay and our methods more generally will be amenable to the study of other neuronal circuits and behaviours in freely behaving flies and in other organisms.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 27 April; accepted 12 July 2016.
Published online 1 August 2016.

1. Reppert, S. M. & Weaver, D. R. Coordination of circadian timing in mammals. *Nature* **418,** 935–941 (2002).
2. Moore, R. Y. Organization and function of a central nervous system circadian oscillator: the suprachiasmatic hypothalamic nucleus. *Fed. Proc.* **42,** 2783–2789 (1983).
3. Rieger, D., Shafer, O. T., Tomioka, K. & Helfrich-Förster, C. Functional analysis of circadian pacemaker neurons in *Drosophila melanogaster*. *J. Neurosci.* **26,** 2531–2543 (2006).
4. Shang, Y., Griffith, L. C. & Rosbash, M. Light-arousal and circadian photoreception circuits intersect at the large PDF cells of the *Drosophila* brain. *Proc. Natl Acad. Sci. USA* **105,** 19587–19594 (2008).
5. Cavanaugh, D. J. *et al.* Identification of a circadian output circuit for rest:activity rhythms in *Drosophila*. *Cell* **157,** 689–701 (2014).
6. Agosto, J. *et al.* Modulation of GABA$_A$ receptor desensitization uncouples sleep onset and maintenance in *Drosophila*. *Nat. Neurosci.* **11,** 354–359 (2008).
7. Tataroglu, O. & Emery, P. Studying circadian rhythms in *Drosophila melanogaster*. *Methods* **68,** 140–150 (2014).
8. Renn, S. C., Park, J. H., Rosbash, M., Hall, J. C. & Taghert, P. H. A *pdf* neuropeptide gene mutation and ablation of PDF neurons each cause severe abnormalities of behavioral circadian rhythms in *Drosophila*. *Cell* **99,** 791–802 (1999).
9. Grima, B., Chélot, E., Xia, R. & Rouyer, F. Morning and evening peaks of activity rely on different clock neurons of the *Drosophila* brain. *Nature* **431,** 869–873 (2004).
10. Guo, F., Cerullo, I., Chen, X. & Rosbash, M. PDF neuron firing phase-shifts key circadian activity neurons in *Drosophila*. *eLife* **3,** e02780 (2014).
11. Yao, Z. & Shafer, O. T. The *Drosophila* circadian clock is a variably coupled network of multiple peptidergic units. *Science* **343,** 1516–1520 (2014).
12. Stoleru, D., Peng, Y., Agosto, J. & Rosbash, M. Coupled oscillators control morning and evening locomotor behaviour of *Drosophila*. *Nature* **431,** 862–868 (2004).
13. Kunst, M. *et al.* Calcitonin gene-related peptide neurons mediate sleep-specific circadian output in *Drosophila*. *Curr. Biol.* **24,** 2652–2664 (2014).
14. Gilestro, G. F. & Cirelli, C. pySolo: a complete suite for sleep analysis in *Drosophila*. *Bioinformatics* **25,** 1466–1467 (2009).
15. Donelson, N. C. *et al.* High-resolution positional tracking for long-term analysis of *Drosophila* sleep and locomotion using the "tracker" program. *PLoS One* **7,** e37250 (2012).
16. Garbe, D. S. *et al.* Context-specific comparison of sleep acquisition systems in *Drosophila*. *Biol. Open* **4,** 1558–1568 (2015).
17. Pfeiffer, B. D. *et al.* Tools for neuroanatomy and neurogenetics in *Drosophila*. *Proc. Natl Acad. Sci. USA* **105,** 9715–9720 (2008).
18. Zhang, L. *et al.* DN1(p) circadian neurons coordinate acute light and PDF inputs to produce robust daily behavior in *Drosophila*. *Curr. Biol.* **20,** 591–599 (2010).
19. Zhang, Y., Liu, Y., Bilodeau-Wentworth, D., Hardin, P. E. & Emery, P. Light and temperature control the contribution of specific DN1 neurons to *Drosophila* circadian behavior. *Curr. Biol.* **20,** 600–605 (2010).
20. Jenett, A. *et al.* A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Reports* **2,** 991–1001 (2012).
21. Flourakis, M. *et al.* A conserved bicycle model for circadian clock control of membrane excitability. *Cell* **162,** 836–848 (2015).
22. Seluzicki, A. *et al.* Dual PDF signaling pathways reset clocks via TIMELESS and acutely excite target neurons to control circadian behavior. *PLoS Biol.* **12,** e1001810 (2014).
23. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nat. Methods* **11,** 338–346 (2014).
24. Aso, Y. *et al.* Mushroom body output neurons encode valence and guide memory-based action selection in *Drosophila*. *eLife* **3,** e04580 (2014).
25. Inagaki, H. K. *et al.* Optogenetic control of *Drosophila* using a red-shifted channelrhodopsin reveals experience-dependent influences on courtship. *Nat. Methods* **11,** 325–332 (2014).
26. Gradinaru, V. *et al.* Molecular and cellular approaches for diversifying and extending optogenetics. *Cell* **141,** 154–165 (2010).
27. Rieger, D. *et al.* The fruit fly *Drosophila melanogaster* favors dim light and times its activity peaks to early dawn and late dusk. *J. Biol. Rhythms* **22,** 387–399 (2007).
28. Feinberg, E. H. *et al.* GFP Reconstitution Across Synaptic Partners (GRASP) defines cell contacts and synapses in living nervous systems. *Neuron* **57,** 353–363 (2008).
29. Yao, Z., Macara, A. M., Lelito, K. R., Minosyan, T. Y. & Shafer, O. T. Analysis of functional neuronal connectivity in the *Drosophila* brain. *J. Neurophysiol.* **108,** 684–696 (2012).
30. Chen, T. W. *et al.* Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature* **499,** 295–300 (2013).
31. Hamasaka, Y. *et al.* Glutamate and its metabotropic receptor in *Drosophila* clock neuron circuits. *J. Comp. Neurol.* **505,** 32–45 (2007).
32. Diao, F. *et al.* Plug-and-play genetic access to *Drosophila* cell types using exchangeable exon cassettes. *Cell Reports* **10,** 1410–1421 (2015).
33. Liu, W. W. & Wilson, R. I. Glutamate is an inhibitory neurotransmitter in the *Drosophila* olfactory system. *Proc. Natl Acad. Sci. USA* **110,** 10294–10299 (2013).
34. Abruzzi, K., Chen, X., Nagoshi, E., Zadina, A. & Rosbash, M. RNA-seq profiling of small numbers of *Drosophila* neurons. *Methods Enzymol.* **551,** 369–386 (2015).
35. Collins, B. *et al.* Differentially timed extracellular signals synchronize pacemaker neuron clocks. *PLoS Biol.* **12,** e1001959 (2014).
36. Collins, B., Kane, E. A., Reeves, D. C., Akabas, M. H. & Blau, J. Balance of activity between LN$_v$s and glutamatergic dorsal clock neurons promotes robust circadian rhythms in *Drosophila*. *Neuron* **74,** 706–718 (2012).
37. Helfrich-Förster, C. Differential control of morning and evening components in the activity rhythm of *Drosophila melanogaster*—sex-specific differences suggest a different quality of activity. *J. Biol. Rhythms* **15,** 135–154 (2000).
38. Low, K. H., Lim, C., Ko, H. W. & Edery, I. Natural variation in the splice site strength of a clock gene and species-specific thermal adaptation. *Neuron* **60,** 1054–1067 (2008).
39. Cao, W. & Edery, I. A novel pathway for sensory-mediated arousal involves splicing of an intron in the period clock gene. *Sleep* **38,** 41–51 (2015).

40. Majercak, J., Sidote, D., Hardin, P. E. & Edery, I. How a circadian clock adapts to seasonal decreases in temperature and day length. *Neuron* **24,** 219–230 (1999).
41. Masuyama, K., Zhang, Y., Rao, Y. & Wang, J. W. Mapping neural circuits with activity-dependent nuclear import of a transcription factor. *J. Neurogenet.* **26,** 89–102 (2012).
42. Stanewsky, R. *et al.* The *cry^b* mutation identifies cryptochrome as a circadian photoreceptor in *Drosophila*. *Cell* **95,** 681–692 (1998).
43. Hardin, P. E., Hall, J. C. & Rosbash, M. Circadian oscillations in period gene mRNA levels are transcriptionally regulated. *Proc. Natl Acad. Sci. USA* **89,** 11711–11715 (1992).
44. Benito, J., Zheng, H., Ng, F. S. & Hardin, P. E. Transcriptional feedback loop regulation, function, and ontogeny in *Drosophila*. *Cold Spring Harb. Symp. Quant. Biol.* **72,** 437–444 (2007).
45. Picot, M., Klarsfeld, A., Chélot, E., Malpel, S. & Rouyer, F. A role for blind DN2 clock neurons in temperature entrainment of the *Drosophila* larval brain. *J. Neurosci.* **29,** 8312–8320 (2009).
46. Prober, D. A., Rihel, J., Onah, A. A., Sung, R. J. & Schier, A. F. Hypocretin/orexin overexpression induces an insomnia-like phenotype in zebrafish. *J. Neurosci.* **26,** 13400–13410 (2006).
47. Liang, X., Holy, T. E. & Taghert, P. H. Synchronous Drosophila circadian pacemakers display nonsynchronous Ca^{2+} rhythms *in vivo*. *Science* **351,** 976–981 (2016).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank M. Diaz, N. Nguyen, R. Spann and K. Kerr for generous help, and X. Gao and L. Luo for sending us *LexAop-LUC* flies, and O. Shafer and A. Sehgal for helpful discussion and comments on early versions of this manuscript. This work was supported in part by NIH R01 MH067284 (LCG).

Author Contributions F.G. and M.R. conceived and designed the experiments. F.G. performed behavioural experiments. F.G. and J.Y. performed immunocytochemistry and imaging experiments. F.G. and H.J.J. setup the recording system. K.C.A. and W.L. performed and quantified the mRNA profiling data. F.G. and J.Y. analysed data. F.G., L.G. and M.R. prepared the figures and wrote the paper, with feedback from all authors.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R. (rosbash@brandeis.edu).

Reviewer Information *Nature* thanks A. Sehgal and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Data reporting. No statistical methods were used to determine sample size. The experiments were not randomized and the investigators were not blinded during experiments and outcome assessment.

Fly strains. *DvPdf-GAL4* was provided by J. H. Park; *Clk4.1M-GAL4* was from P. Hardin; *UAS-dTrpA1* (2nd) was from P. Garrity; *UAS-CaLexA* was from J. Wang[41]; *UAS-TNT* and *UAS-Tet* were from H. Amrein; *Pdf-GAL80* and *CRY-GAL80* are described by Stoleru *et al.*[12]; *LexA-P2X2* and *Clk856-GAL4* were from O. Shafer[11,29]. *UAS-CD4::spGFP1-10* and *LexAop-CD4::spGFP11* were from K. Scott; *Clk4.1M-lexA* was from A. Sehgal[5], *LexAop-LUC* was generated by X. Gao and L. Luo[48]. *LexAop-dTrpA1* was from G. M. Rubin. *UAS-VGLUT* RNAi 1 (VDRC 104324), UAS-*mGluRA* RNAi 1 (VDRC 103736), *UAS-mGluRA* RNAi 2 (VDRC 1793) were from the Vienna *Drosophila* Resource Center (VDRC). The following lines were ordered from the Bloomington Stock Center: *Pdfr (R18H11)-GAL4* (48832), *Pdfr (R18H11)-LexA (52535),UAS-CsChrimson* (55136), *UAS-eNPHR3.0* (36350), *UAS-Denmark* (33064), *UAS-ArcLight* (51056), *UAS-GCaMP6f* (42747),*UAS-syt-GFP* (33064), *UAS-VGLUT* RNAi 2 (40845, 40927), *VGlut^{MI04979}-GAL4* (60312). Flies were reared on standard cornmeal/agar medium supplemented with yeast. The adult flies were entrained in 12:12 light-dark cycles at 25 °C. The flies carrying *GAL4* and *UAS-dTrpA1* were maintained at 21 °C to inhibit dTrpA1 activity.

Locomotor activity. Locomotor activity of individual male flies (aged 3–7 days) was measured with Trikinetics *Drosophila* Activity Monitors or video recording system under 12:12 light:dark conditions. The activity and sleep analysis was performed with a signal-processing toolbox implemented in MATLAB (MathWorks). Group activity was also generated and analysed with MATLAB. For dTrpA1-induced neuronal firing experiments (Fig. 3 and Extended Data Fig. 9), flies were entrained in light:dark for 3–4 days at 21 °C, transferred to 27 °C for two days, followed by 2 subsequent days at 21 °C. The evening activity index (Extended Data Fig. 9) was calculated by dividing the average activity from ZT8–12 by the average activity from ZT 0–12. The behaviour experiments involving RNAi expression (Extended Data Fig. 10b) were done at 27 °C to enhance knockdown efficiency.

Statistical analyses. All statistical analyses were conducted using IBM SPSS software. The sample size was chosen based on the pilot studies to ensure >80% statistical power to detect significant difference between different groups. Animals within the same genotype were randomly allocated to experimental groups and then processed. We were not blind to the group allocation as the experimental design required specific genotypes for experimental and control groups. However, the data analyser was blinded when assessing the outcome. The Wilks–Shapiro test was used to determine normality of data. Normally distributed data were analysed with two-tailed, unpaired Student's *t*-tests, one-way analysis of variance (ANOVA) followed by a Tukey–Kramer HSD test as the post-hoc test or two-way analysis of variance (ANOVA) with post-hoc Bonferroni multiple comparisons. Nonparametrically distributed data were assessed using the Kruskal–Wallis test. Data were presented as mean behavioural responses, and error bars represent the standard error of the mean (s.e.m.). Differences between groups were considered significant if the probability of error was less than 0.05 ($P < 0.05$). Experiments were repeated at least three times and representative data was shown in figures.

Arousal thresholds assay. For mechanical stimulation, individual flies from different groups were loaded into 96-well plates and placed close to a small push–pull solenoid. The tap frequency of the solenoid was directly driven by an Arduino UNO board (Smart Projects). One tap was used as a modest stimulus and ten taps (1 Hz) was used as a strong stimulus. Arousal threshold was measured during the middle of the day (ZT6) and evening (ZT10) with different intensities. The movement of flies before and after the stimulus was monitored by the web camera and the recording videos (1fps) were processed by the MTrack2 plugin in Fiji ImageJ software to convert the videos into binary images and to calculate the trajectory and moving area as well as the percentage of aroused flies.

Feeding of retinal. All trans-retinal (ATR) powder (Sigma) was dissolved in alcohol to prepare a 100 mM stock solution for CsChrimson experiments[23]. 100 μl of this stock solution was diluted in 25 ml of 5% sucrose and 1% agar medium to prepare 400 μM of ATR food. Newly eclosed flies were transferred to ATR food for at least 2 days before optogenetic experiments.

Optogenetics and video recording system. The behavioural setup for the optogenetics and video recording system is schematized in Supplementary Fig. 1. Briefly, flies were loaded into white 96-well Microfluor 2 plates (Fisher) containing 5% sucrose and 1% agar food with or without 400 μM ATR. Back lighting for night vision was supplied by an 850 nm LED board (LUXEON) located under the plate. Two sets of high power LEDs (627 nm) mounted on heat sinks (four LEDs per heat sink) were symmetrically placed above the plate to provide light stimulation. The angle and height of the LEDs were adjusted to ensure uniform illumination. The voltage and frequency of red light pulses were controlled by an Arduino UNO board (Smart Projects). The whole circuit is described in ref. 25. The flat surface and compact wells of the 96-well plate allow uniform illumination, which was difficult to achieve in Trikinetics tubes. We used 627 nm red light pulses at 10 Hz (0.08 mW mm^{-2}) to irradiate flies expressing the red-shifted channelrhodopsin CsChrimson within the DN1s[23]. (The CsChrimson illumination protocol had no effect on halorhodopsin eNpHR3.0). Fly behaviour was recorded by a web camera (Logistic C910) without an infrared filter. We used time-lapse software to capture snapshots at 10 s intervals. The light:dark cycle and temperature was controlled by the incubator, and the light intensity was maintained in a region that allowed entrainment of flies without activating CsChrimson. Fly movement was calculated by Pysolo software and transformed into a MATLAB readable file[14]. 5 pixels per second (50% of the Full Body Length) was defined as a minimum movement threshold[15,16]. The activity and sleep analyses were performed with a signal-processing toolbox implemented in MATLAB (MathWorks) as described above. The design of the invention has been filed for patent.

***In vivo* luciferase assays.** To monitor bioluminescence activity in living flies, we used previously described protocols[49]. White 96-well Microfluor 2 plates (Fisher) were loaded with 5% sucrose and 1% agar food containing 20 mM D-luciferin potassium salt (GOLDBIO). 250 μl of food was added to each well. Individual male or female flies expressing CaLexA–LUC were first anaesthetized with CO_2 and then transferred to the wells. We used an adhesive transparent seal (TopSeal-A PLUS, Perkin Elmer) to cover the plate and poked 2–3 holes in the seal over each well for air exchange. Plates were loaded into the stacker of a TopCount NXT luminescence counter (Perkin Elmer). Assays were carried out in an incubator under light:dark conditions. Luminescence counts were collected for 5–7 days. For temperature shift experiments (Fig. 4b), the incubator temperature was set to 21 °C for 3 days and then increased to 30 °C at ZT 0 of the 4th day. Other experiments were performed at 25 °C. Three different modes were used in our experiments: (1) To record CaLexA–LUC activity only, 9 plates were placed in a stacker, and each plate was sequentially transferred to the TopCount machine for luminescence reading. Every cycle took about 1 h, and the recording was continued for several days. (2) To combine optogenetic stimulation with the luciferase assays (Fig. 4a and Extended Data Fig. 6a), we replaced the stacker with a chamber of our own design (Fig. 4a). 627 nm LEDs mounted to a pair of heat sinks were symmetrically positioned in the chamber to ensure uniform illumination of the 96-well plate (0.08 mW mm^{-2} for CsChrimson stimulation and 1 mW mm^{-2} for eNPHR3.0 stimulation). Flies pre-fed with ATR were loaded into a plate. Single plates stayed in the LED chamber for 8 min and then automatically transferred to the TopCount for luminescence reading for 2 min. (3) To assay fly movement in 96-well plates and CaLexA–LUC activity at the same time, single plates were recorded using a web camera attached to the top of chamber (Fig. 4a). During each hour, the plate sat in the video chamber for 58 min and then was automatically transferred to the TopCount machine for a 2 min luminescence reading. The raw data were analysed in MATLAB and in Microsoft Excel. All experiments were repeated at least three times.

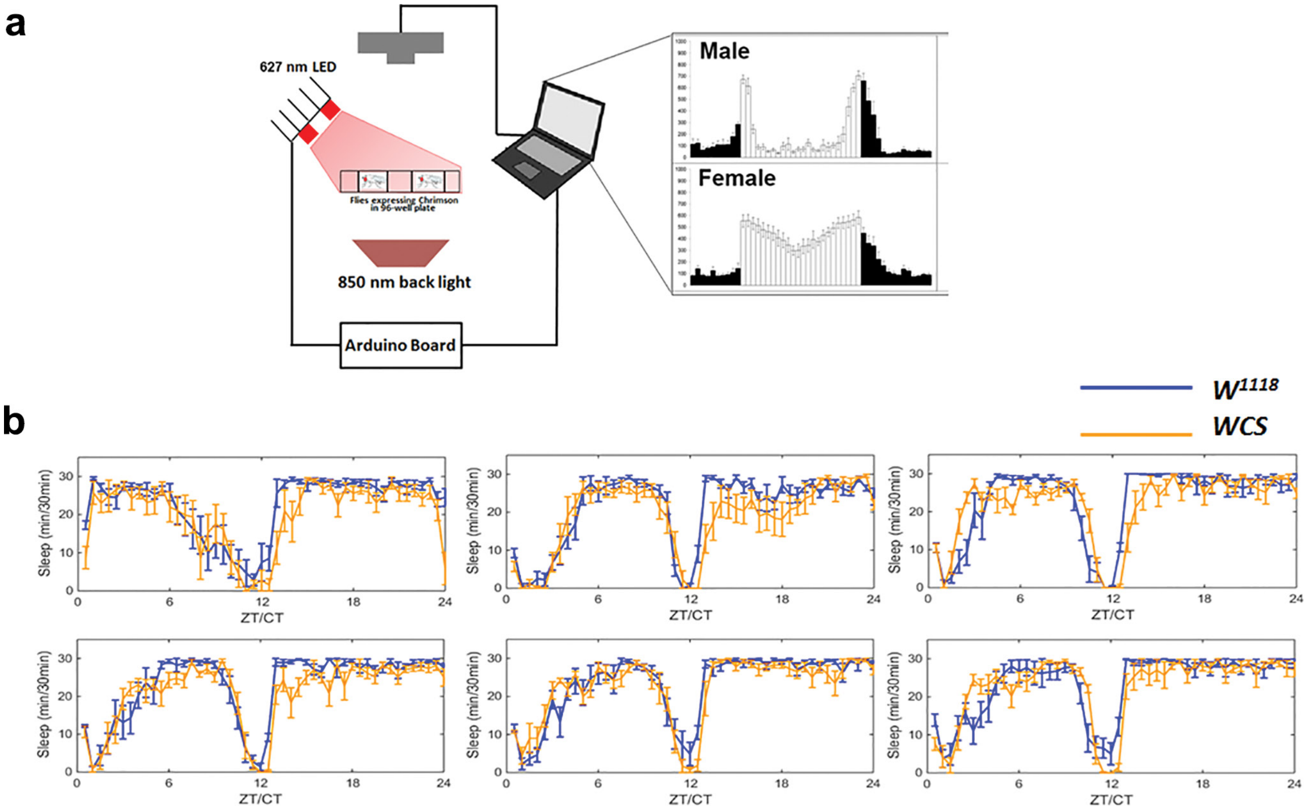
Fly brain immunocytochemistry. Immunostaining was performed as described[50]. Fly heads were removed and fixed in PBS with 4% paraformaldehyde and 0.008% Triton X-100 for 45–50 min at 4 °C. Fixed heads were washed in PBS with 0.5% Triton X-100 and dissected in PBS. The brains were blocked in 10% goat serum (Jackson Immunoresearch) and subsequently incubated with primary antibodies at 4 °C overnight or longer. For VGLUT and GFP co-staining, a rabbit anti-DVGlut (1:10,000) and a mouse anti-GFP antibody (Invitrogen; 1:1,000) antibody were used as primary antibodies. For GRASP staining, a mouse anti-GFP monoclonal antibody (Invitrogen; 1:1,000) and a rabbit anti-GFP antibody (Roche; 1:200) were used. After washing with 0.5% PBST three times, the brains were incubated with Alexa Fluor 633 conjugated anti-rabbit and Alexa Fluor 488 conjugated anti-mouse (Molecular Probes) at 1:500 dilution. The brains were washed three more times before being mounted in Vectashield Mounting Medium (Vector Laboratories) and viewed sequentially in 1.1 μm sections on a Leica SP5 confocal microscope. To compare the fluorescence signals from different conditions, the laser intensity and other settings were set at the same level during each experiment. Fluorescence signals were quantified by ImageJ as described.

mRNA profiling from E cells and DN1s. mRNA profiling from E cells and DN1s was performed as previously described[34]. DN1s and E cells were purified from *Clk4.1M-GAL4, UAS-EGFP* flies (DN1s) and *Dv-Pdf-GAL4, UAS-EGFP, PDF-RFP* flies, (E cells; GFP^{+}RFP^{-} cells), respectively. Flies were entrained for 3 days and then collected every 4 h for a total of six time points. Two replicates of six time points were performed for each cell type. Sequencing data were aligned to the *Drosophila* genome using TopHat[51]. Gene expression was quantified using the End Sequencing Analysis Toolkit (ESAT; publicly available at http://garberlab.umassmed.edu/software/esat/). ESAT quantifies gene expression only using information from the 3′-end of the genes.

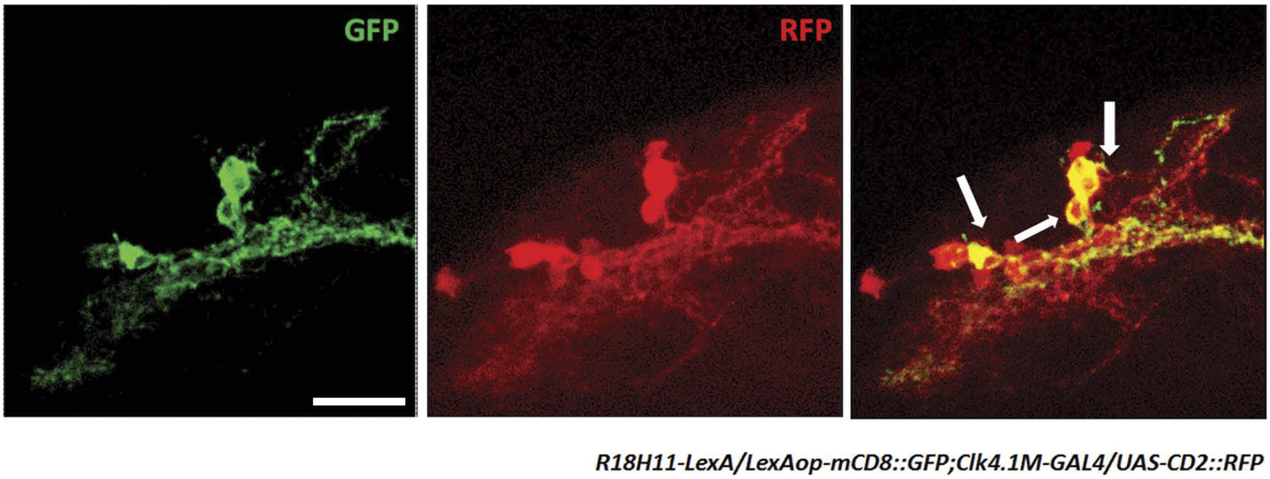
Functional fluorescence imaging. Imaging experiments were performed as previous described[52]. Adult male fly brains were dissected in ice-cold haemolymph-like saline (AHL) (108 mM NaCl, 5 mM KCl, 2 mM CaCl2, 8.2 mM $MgCl_2$, 4 mM $NaHCO_3$, 1 mM NaH_2PO_4-H_2O, 5 mM trehalose, 10 mM sucrose, 5 mM HEPES; pH 7.5). Brains were then pinned to a layer of Sylgard (Dow Corning) silicone under a small bath of AHL contained within a recording/perfusion chamber (Warner Instruments) and bathed with room temperature AHL. Brains expressing GCaMP6f and Arclight were exposed to fluorescent light for approximately 30 s before imaging to allow for baseline fluorescence stabilization. Perfusion flow was established over the brain with a gravity-fed ValveLink perfusion system (Automate Scientific). ATP or glutamate was delivered by switching the perfusion flow from the main AHL line to another channel containing diluted compound after 30 s of baseline recording for the desired durations followed by a return to AHL flow. For the mGluRA antagonist imaging experiments, 700 nM LY341495 (Tocris Bioscience) was used to block the glutamate-induced inhibition. Imaging was performed using an Olympus BX51WI fluorescence microscope (Olympus) under an Olympus ×40 (0.80 W, LUMPlanFl) or ×60 (0.90W, LUMPlanFl) water-immersion objective, and all recordings were captured using a charge-coupled device camera (Hamamatsu ORCA C472-80-12AG). For GCaMP6f and Arclight imaging, the following filter sets were used (Chroma Technology): excitation, HQ470/×40; dichroic, Q495LP; emission, HQ525/50 m. Frames were captured at 2 Hz with 4 × binning for either 2 min or 4 min using μManager acquisition software[52]. Neutral density filters (Chroma Technology) were used for all experiments to reduce light intensity and to limit photobleaching.

For recordings using GCaMP6f and Arclight, ROIs were analysed using custom software developed in ImageJ[52] and National Institute of Health. The fluorescence change was calculated by using the formula: $\Delta F/F = (F_n - F_0)/F_0 \times 100\%$, where F_n is the fluorescence at time point n, and F_0 is the fluorescence at time 0. The fluorescence was calibrated by subtracting the background fluorescence value. To compare the fluorescence change between neurons in the same brain, fluorescence activities from different neurons were normalized to the highest fluorescence level during the recording time window.

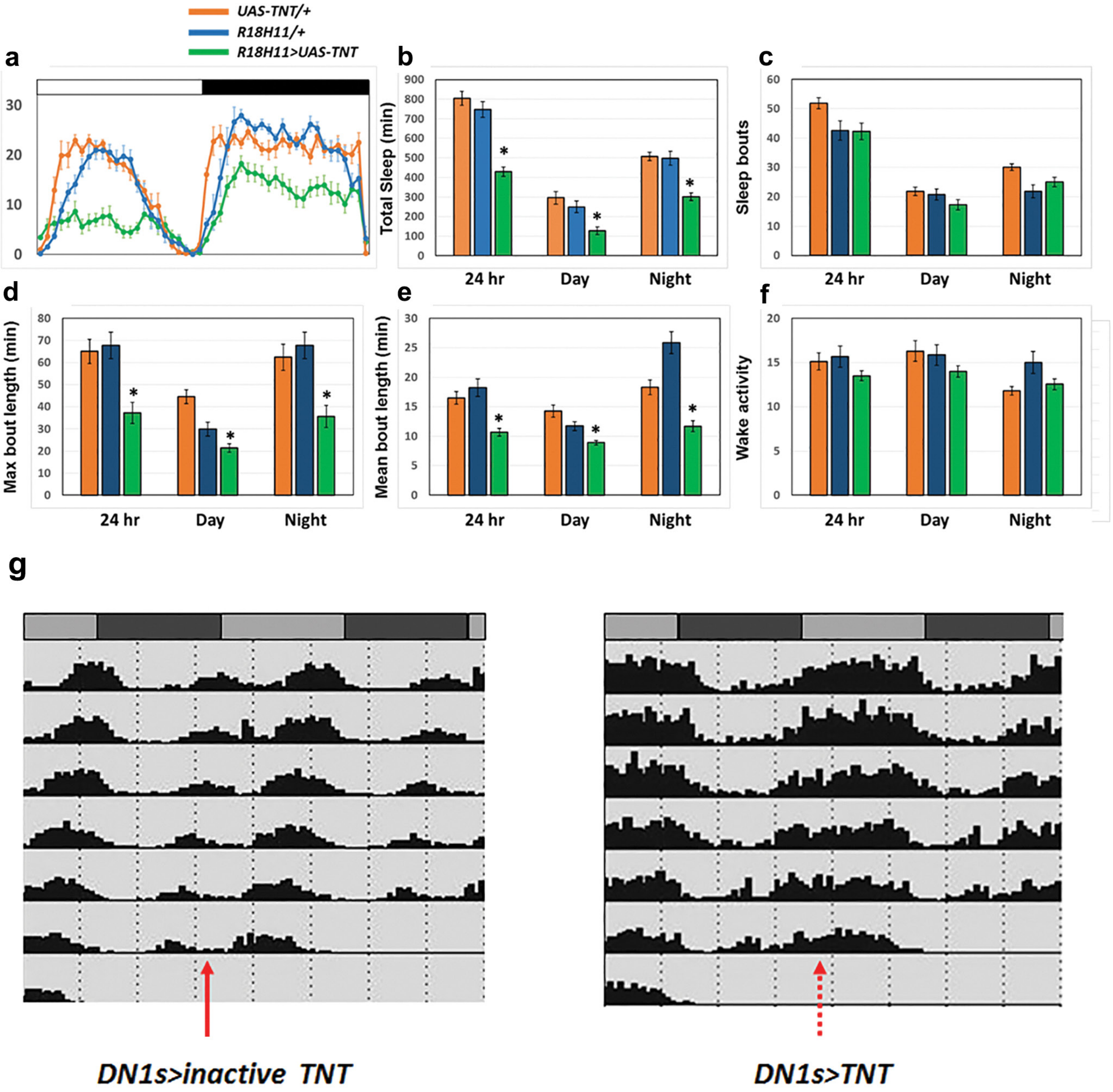
48. Gao, X. J. *et al.* A transcriptional reporter of intracellular Ca^{2+} in *Drosophila*. *Nat. Neurosci.* **18,** 917–925 (2015).
49. Stanewsky, R. Analysis of rhythmic gene expression in adult *Drosophila* using the firefly luciferase reporter gene. *Methods Mol. Biol.* **362,** 131–142 (2007).
50. Tang, C. H., Hinteregger, E., Shang, Y. & Rosbash, M. Light-mediated TIM degradation within *Drosophila* pacemaker neurons (s-LNvs) is neither necessary nor sufficient for delay zone phase shifts. *Neuron* **66,** 378–385 (2010).
51. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25,** 1105–1111 (2009).
52. Haynes, P. R., Christmann, B. L. & Griffith, L. C. A single pair of neurons links sleep to memory consolidation in *Drosophila melanogaster*. *eLife* **4,** e03868 (2015).



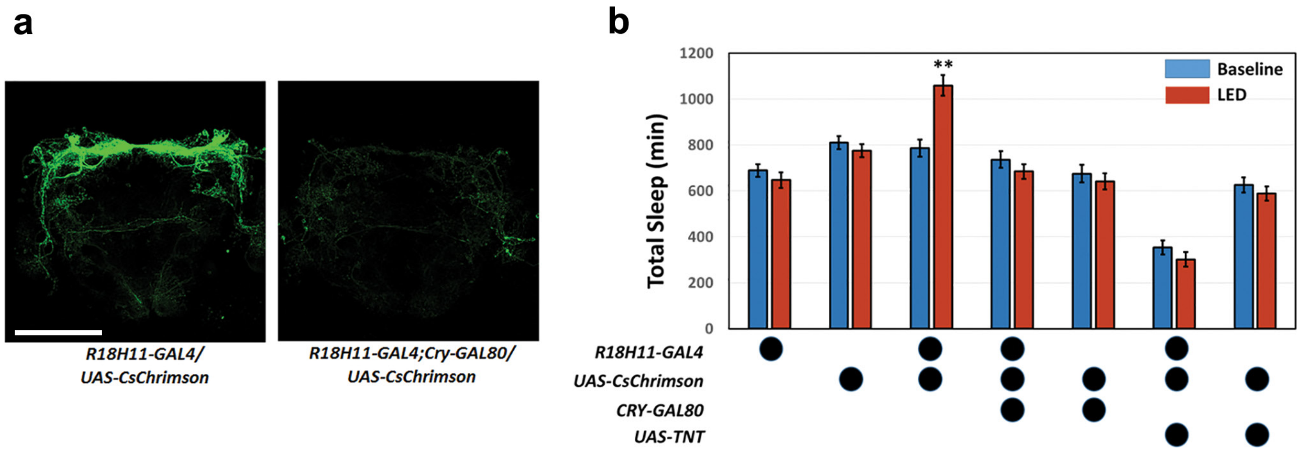
Extended Data Figure 1 | Schematics of video recording and optogenetic strategies. a, Flies expressing CsChrimson were placed in 96-well plates and video recorded with a camera without an infrared filter (left). An 850 nm infrared back light provides illumination for recording in both light and dark periods. A set of 627 nm LEDs was carefully positioned and combined with a diffuser to ensure uniform irradiation for stimulation. The voltage and pulse frequency were controlled by an Arduino UNO board as described in Methods. Representative data from a video recording of male and female activity (right) in light:dark are shown. **b**, Sleep data of two control genotypes in 96-well plate mode for 6 days. Results with error bars are mean ± s.e.m. *n* = 15–16 for each group.



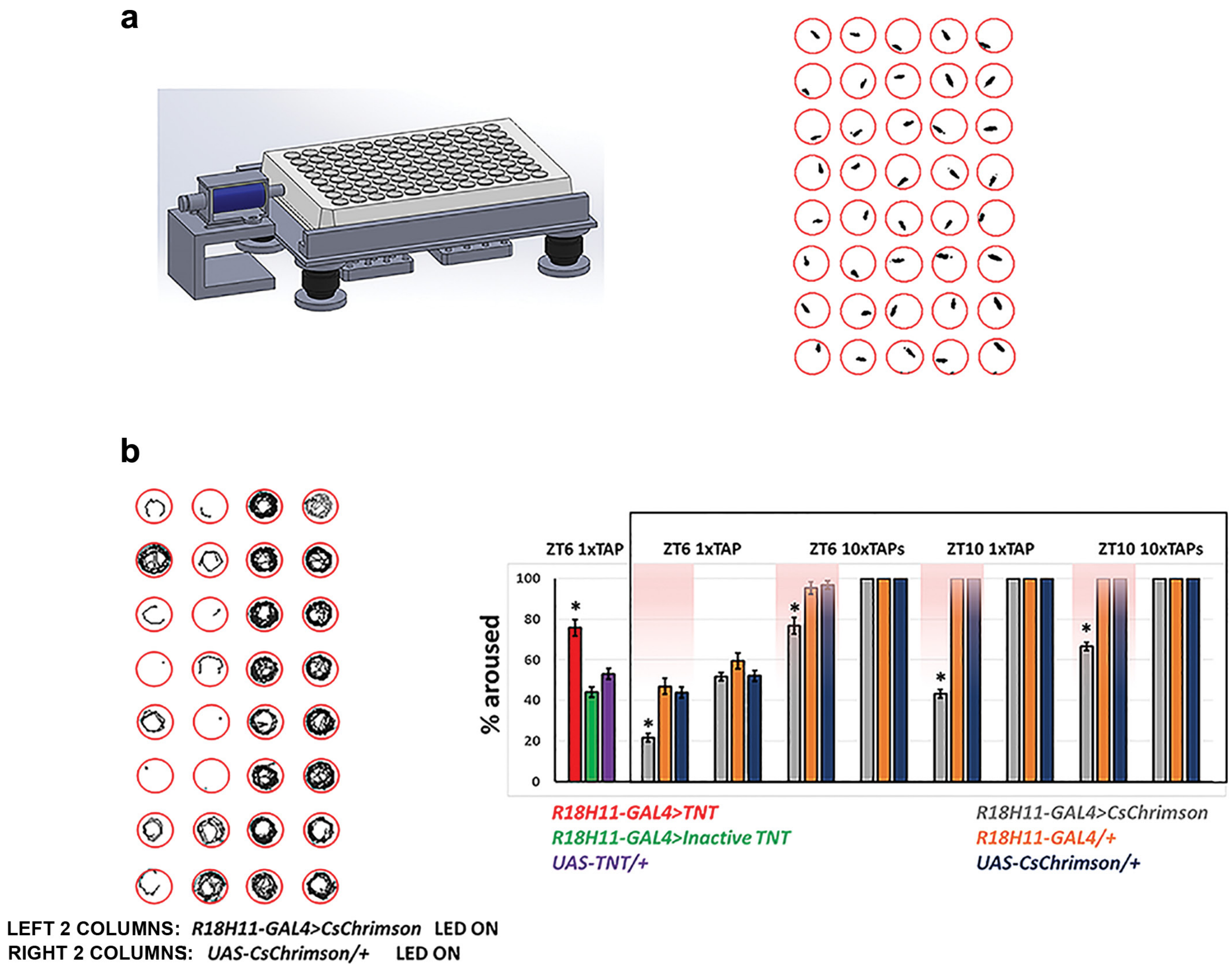
Extended Data Figure 2 | The *R18H11* driver labels a subgroup of CLK4.1M-defined DN1s. Confocal stack of images showing antibody staining for GFP (left) and RFP (middle) and the overlay (right) in the dorsal brain of *R18H11-LexA/LexAop-mCD8::GFP;Clk4.1M-GAL4/UAS-rCD2::RFP* flies. Scale bar, 20 μm.



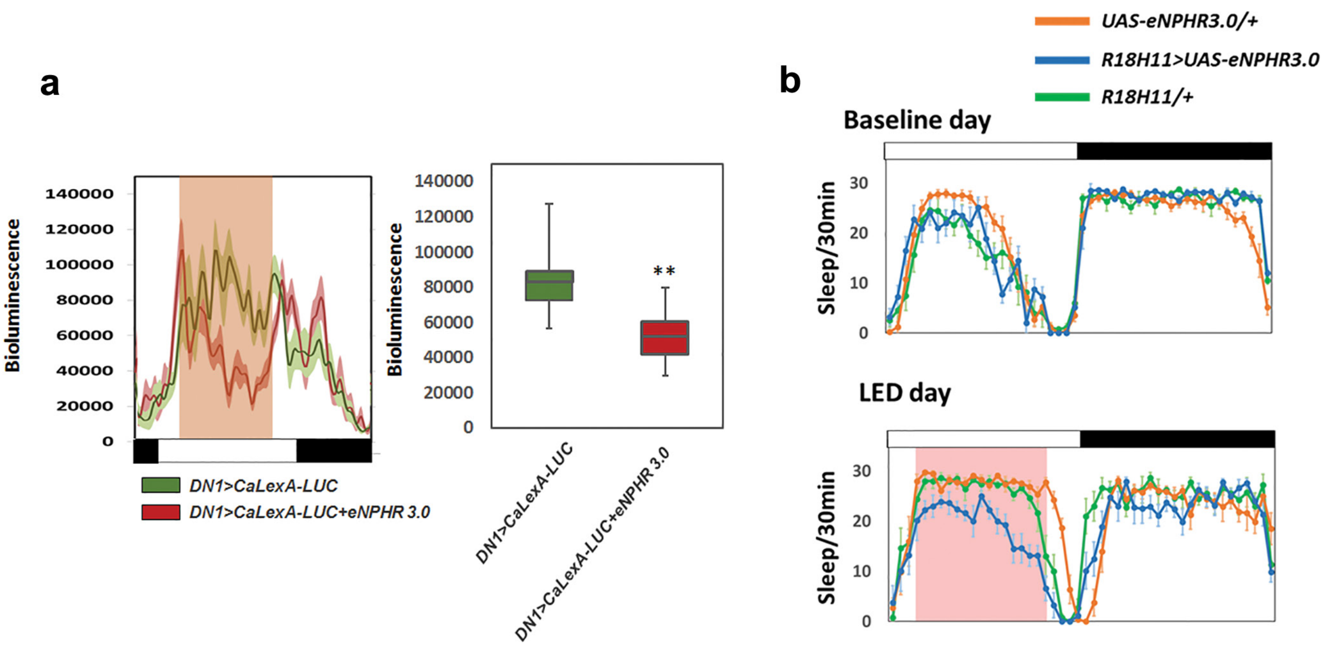
Extended Data Figure 3 | Blocking neurotransmitter output of DN1s affects sleep parameters but not DD rhythmicity. a–f, Total sleep, maximum sleep bout duration and mean sleep bout duration of control groups (*R18H11-GAL4/+*, *UAS-TNT/+*) and the experimental group (*R18H11-GAL4/UAS-TNT*) have significant differences. Results with error bars are mean ± s.e.m. $n = 32$ for each group. One-way ANOVA was performed to detect significant genotype effects for total sleep ($P < 0.0001$), daytime sleep ($P < 0.0001$), night-time sleep ($P < 0.0001$) (**a**, **b**), maximum bout duration ($P = 0.00288$), maximum daytime bout duration ($P < 0.0001$), maximum night-time bout duration ($P = 0.000388$) (**d**), mean bout duration ($P < 0.0001$), mean daytime bout duration ($P < 0.0001$), mean night-time bout duration ($P < 0.0001$) (**e**). Asterisks denote significant differences from parental controls in Tukey's post-hoc test ($P < 0.01$). **g**, Neurotransmitter release from DN1s is not required for DD rhythmicity. Locomotor behaviour of control (*Clk4.1M-GAL4/UAS-Tet*) and experimental (*Clk4.1M-GAL4/UAS-TNT*) male flies was monitored for 6 days in DD. Both control (left) and experimental (right) flies maintained strong rhythmicity. Note that the experimental group showed much less daytime sleep and higher activity level (dashed red arrow-right panel) than the control group (solid red arrow-left panel). $n = 32$ for each group.



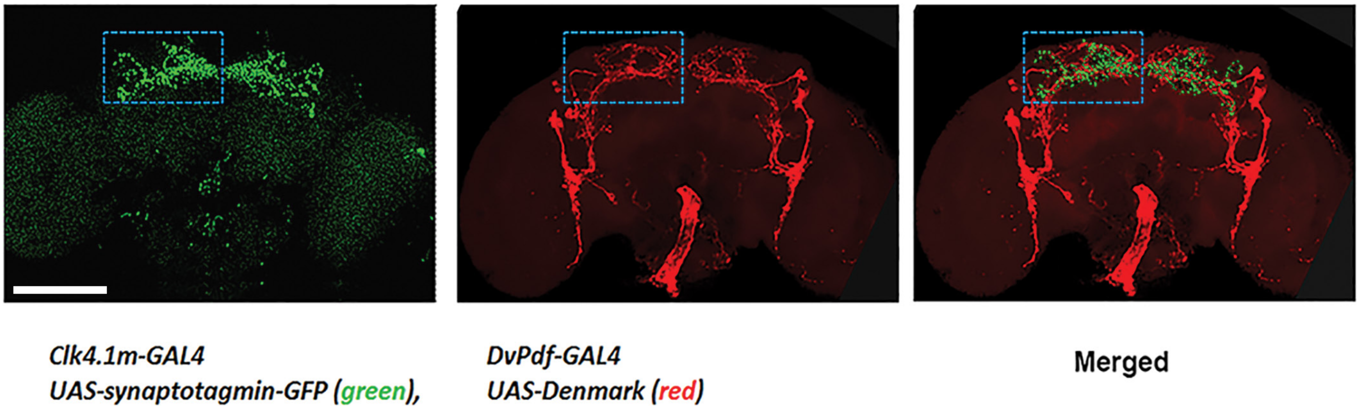
Extended Data Figure 4 | Co-expression of *CRY-GAL80* or *TNT* blocks the sleep-promoting effect of DN1 activation. a, *R18H11-GAL4/UAS-CsChrimson* (left) and *R18H11-GAL4, CRY-GAL80/UAS-CsChrimson* (right) brains were dissected and stained with anti-GFP (green). Scale bar, 100 µm. **b**, Comparison of total sleep in the baseline day (blue) to total sleep during a 24 h LED stimulation day (red) for each genotype. $n = 32$ for *R18H11-GAL4/UAS-CsChrimson* and $n = 24$ for the other groups. Results with error bars are mean ± s.e.m. $^{**}P < 0.001$ by post-hoc Bonferonni multiple comparisons. Two-way ANOVA was performed to detect a significant LED stimulation effect ($P = 0.00227674$), a genotype effect ($P < 0.0001$) and interaction ($P < 0.0001$).



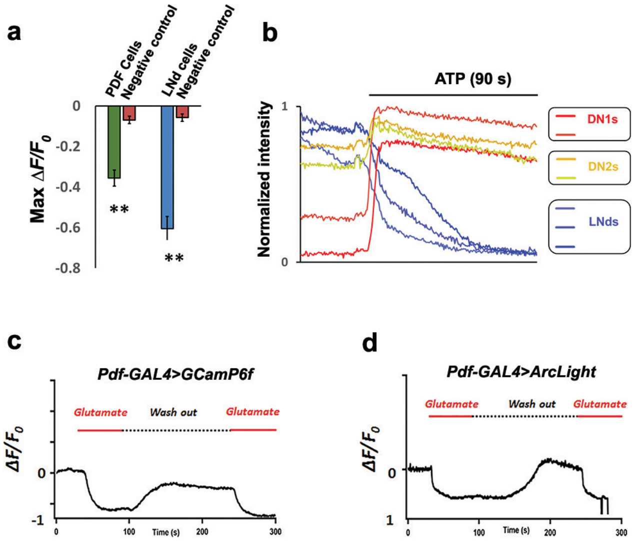
Extended Data Figure 5 | Arousal threshold is affected by manipulation of DN1 activity. a, Mechanical stimulation setup for measuring arousal threshold. The set-up is illustrated on the left. A 96-well plate is loaded onto the device and a small push–pull solenoid is positioned on the side of the plate. The solenoid can be programed to tap the plate at different frequencies and times. A web camera monitors fly movement in the wells, and the video is analysed with Fiji ImageJ software to track flies. An example image is shown on the right. **b**, Activation of DN1s increases arousal threshold. The left panel shows representative trajectories (5 min traces) of experimental and control flies after a strong (10 tap) stimulus at ZT6 when the LED is on. The right panel shows the percentages of flies for the indicated genotypes that transitioned from immobility to an active state in response to the stimulus. The pink background indicates LED stimulation. $n = 24$ for *R18H11-GAL4/UAS-CsChrimson* and *R18H11-GAL4/UAS-TNT* groups, $n = 16$ for parental control groups. Results with error bars are mean ± s.e.m. One-way ANOVA was performed to detect significant genotype effects for arousal levels of *R18H11>CsChrimson* LED group ($P < 0.0001$) and *R18H11>TNT* group ($P < 0.0001$). Asterisks denote significant differences from parental controls in Tukey's post-hoc test ($P < 0.01$).



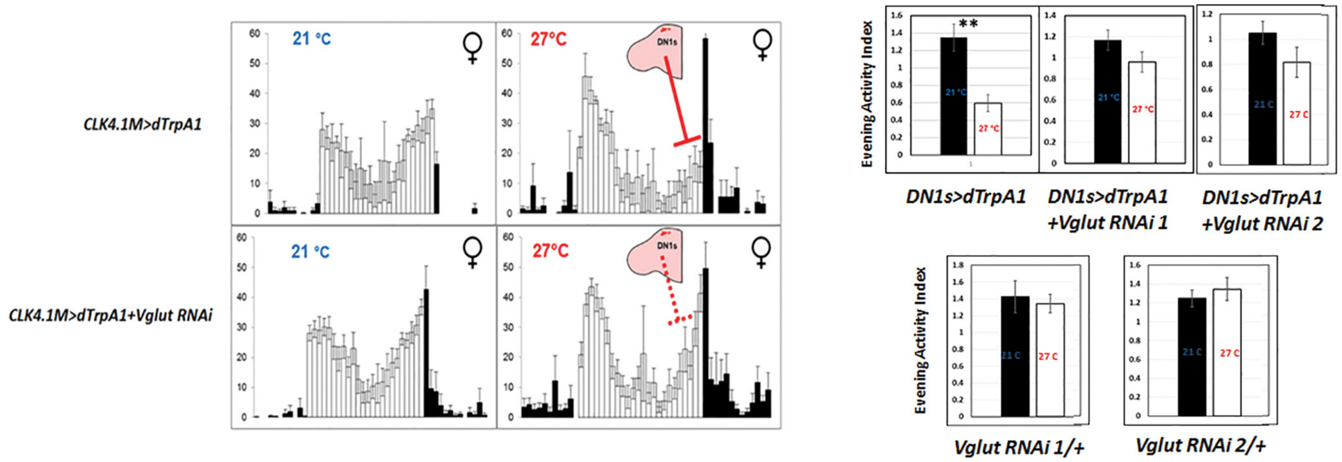
Extended Data Figure 6 | e*NPHR3.0* blocks DN1 neuronal activity and decreases the siesta. a, The luminescence traces from the CaLexA–LUC sensor reflect neuronal activity after LED-induced *eNPHR3.0* inhibition from ZT 2.5 to ZT 9.5. The mean LUC activity level (arbitrary units) from control and experimental groups is quantified on the right. Results with shading are mean ± s.e.m. Box boundaries represent the first and third quartiles, whiskers are 1.5 interquartile range. $n = 24$ for each group. **b**, Sleep from a baseline day and from a LED stimulation day of *R18H11>UAS-eNPHR3.0*, *UAS-eNPHR3.0/+* and *R18H11/+* flies. Pink background represents LED stimulation. $n = 16$ for each group. Results with error bars are mean ± s.e.m.



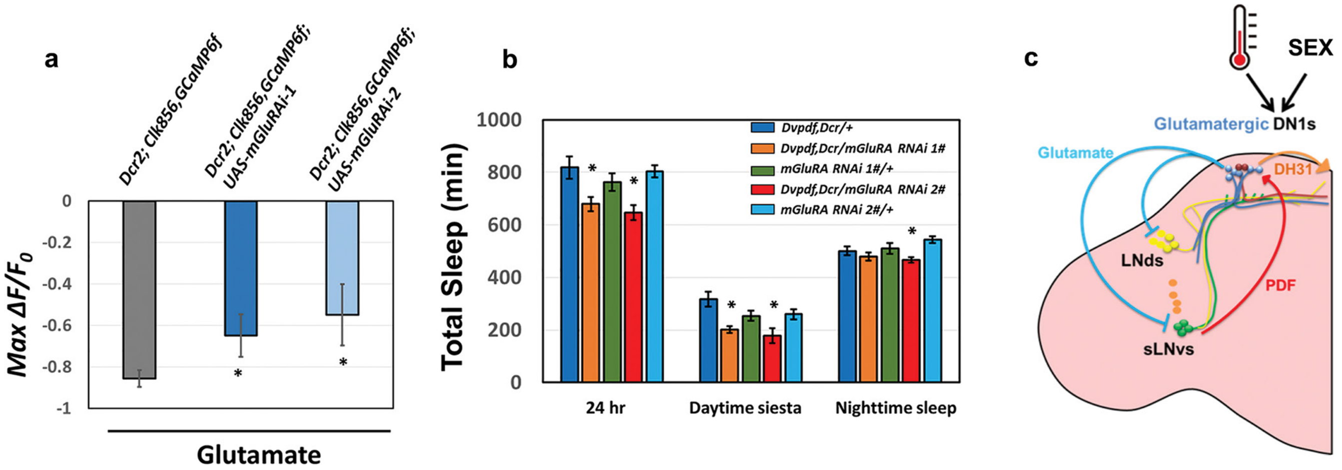
Extended Data Figure 7 | The dendritic region of E cells overlaps with the presynaptic region of DN1s. *Clk4.1M-GAL4>UAS-synaptotagmin-GFP* brains were dissected and stained with anti-GFP antibodies to identify the DN1 presynaptic regions (green; left panel). To identify the dendritic regions of E cells and M cells, *DvPdf-GAL4>UAS-Denmark* brains were dissected and stained with anti-DsRed antibodies (red; middle panel). These patterns were aligned and overlaid (merged panel on the right).



Extended Data Figure 8 | Glutamate reduces calcium levels in PDF neurons and hyperpolarizes their membrane potential. a, Quantification of peak GCaMP6f changes shown in Fig. 2c. Average maximum changes for LN$_v$s and dorsal lateral neurons are shown. $n = 8$ for control group, 7 for PDF cells and 11 for dorsal lateral neurons. $^{**}P < 0.001$ by unpaired two-tailed Student's t-test and results with error bars are mean ± s.e.m. **b**, Normalized calcium traces in different circadian neuron subgroups imaged concurrently in the same representative brain. **c**, **d**, 5 mM glutamate was applied to exposed dissected fly brains (*Pdf-GAL4>GCaMP6f*, (**c**) and *Pdf-GAL4>Arclight* (**d**)) and induced a calcium decrease and hyperpolarized these core pacemakers ($\Delta F/F$ represents the evoked fluorescence change from baseline). The red solid line indicates time of glutamate application. The dashed line indicates time of vehicle application. Panels show representative data of 6 brains.



Extended Data Figure 9 | Decreasing DN1 VGLUT levels blocks the E peak reduction caused by DN1 firing. Reducing VGLUT activity within the DN1s decreases the DN1-activation effect (left). The locomotor activity patterns of *UAS-dTrpA1/+; CLK4.1M-GAL4/+* female flies (upper panel) and *UAS-dTrpA1/+; CLK4.1M-GAL4/UAS-VGlut* RNAi female flies (lower panel) at 21 °C and 27 °C are shown. The colour of the bars indicates either daytime (white) or night-time (black). Evening activity index was calculated as described in the Methods for the indicated genotypes (right). White and black bars represent data from low and high temperature respectively. $^{**}P < 0.001$ by unpaired two-tailed Student's *t*-test. $n = 24$ for each group. Results with error bars are mean ± s.e.m.



Extended Data Figure 10 | Reducing mGluRA expression in pacemaker neurons reduces the inhibitory effect of glutamate as well as the siesta. a, The peak decrease of GCaMP6f in circadian cells after applying glutamate to control and mGluRA knockdown flies. The genotypes are shown above the bars. $n = 6$ *UAS-Dcr2; Clk856-GAL4, UAS-GCaMP6f* and $n = 8$–9 *UAS-Dcr2; Clk856-GAL4, UAS-GCaMP6f; UAS-mGluRA* RNAi groups. *$P < 0.05$ by unpaired *t*-test. **b**, Comparison of total sleep, daytime siesta and night-time sleep in different genotypes. $n = 32$ for each group. *$P < 0.05$ by one-way ANOVA with Tukey's post-hoc test. Results with error bars are mean ± s.e.m (**a**, **b**). **c**, A temporally constrained negative feedback core pacemaker–DN1 circuit regulates the fly activity–sleep pattern. Early in the day, M pacemaker neurons activate the DN1s via the PDF neuropeptide, and DN1s release DH31 to enhance morning arousal. Later in the day, glutamate release from DN1s inhibits M cells and E cells, promotes the siesta, decreases the evening activity peak and initiates night-time sleep. A cycling mRNA that encodes inhibitory glutamate receptors in pacemaker cells may help direct this inhibition to the late day. This feedback circadian circuit shapes the bimodal locomotor activity peak and sleep–wake cycles under normal conditions. The higher daily neuronal activity in male DN1s compared to female DN1s promotes the sexually dimorphic activity/sleep pattern. DN1s also integrate environmental information such as temperature to promote sleep plasticity.

doi:10.1038/nature19069

Defining the clonal dynamics leading to mouse skin tumour initiation

Adriana Sánchez-Danés[1]*, Edouard Hannezo[2,3,4]*, Jean-Christophe Larsimont[1], Mélanie Liagre[1], Khalil Kass Youssef[1], Benjamin D. Simons[2,3,4] & Cédric Blanpain[1,5]

The changes in cell dynamics after oncogenic mutation that lead to the development of tumours are currently unknown. Here, using skin epidermis as a model, we assessed the effect of oncogenic hedgehog signalling in distinct cell populations and their capacity to induce basal cell carcinoma, the most frequent cancer in humans. We found that only stem cells, and not progenitors, initiated tumour formation upon oncogenic hedgehog signalling. This difference was due to the hierarchical organization of tumour growth in oncogene-targeted stem cells, characterized by an increase in symmetric self-renewing divisions and a higher p53-dependent resistance to apoptosis, leading to rapid clonal expansion and progression into invasive tumours. Our work reveals that the capacity of oncogene-targeted cells to induce tumour formation is dependent not only on their long-term survival and expansion, but also on the specific clonal dynamics of the cancer cell of origin.

Cancer arises through the acquisition of oncogenic mutations[1]. How such oncogenic mutations affect the rate of stem and progenitor cell proliferation and the proportion of divisions that result in symmetric and asymmetric fate is currently poorly understood. Recent studies following oncogenic activation in mouse gut before tumour formation showed that intestinal stem cells (SCs) acquire a proliferative advantage over their wild-type neighbours, leading to precocious clonal fixation of mutant crypts[2,3]. However, the question of whether and how mutant crypts expand and progress into invasive tumours remains unknown.

Basal cell carcinoma (BCC) is the most frequently occurring type of tumour in humans, with more than 5 million new cases diagnosed each year worldwide. BCCs arise from the constitutive activation of the hedgehog (HH) pathway through either Patched (Ptch1) loss of function or Smoothened (Smo) gain of function[4]. Different mouse models of BCC using *Ptch1* deletion or oncogenic SmoM2 mutant expression induce the formation of tumours that resemble superficial human BCC[5]. The skin epidermis contains distinct types of SCs that contribute to the homeostasis of discrete regions of epidermis[6]. Interfollicular epidermis (IFE) is maintained by SCs targeted by K14-CreER, that drives the expression of inducible CreER under the control of the Keratin 14 promoter; and committed progenitors (CPs) targeted by Inv-CreER, in which the CreER is expressed under the control of the Involucrin (Inv) promoter in tail, ear, back and ventral skin epidermis[7,8]. Activation of oncogenic HH signalling through SmoM2 expression or *Ptch1* deletion in these different tissues using K14-CreER, which targets both SCs and CPs, induces BCC formation[7,9–12]. However, the question of whether and how SmoM2 expression in SCs and/or CPs drives BCC formation remains unresolved.

SCs but not CPs initiate BCC formation

To determine whether SCs and CPs can induce BCC, we induced oncogenic SmoM2 expression exclusively in CPs using Inv-CreER, and in both CPs and SCs using K14-CreER[7] at the same clonal density (Fig. 1a and Extended Data Fig. 1a). As previously reported, activation of SmoM2 expression using K14-CreER induced BCC, characterized by invasion into the dermis and branched morphology, in both tail and ear epidermis[9–11] (Fig. 1b). In sharp contrast, activation of SmoM2 expression in CPs using Inv-CreER lead to pre-neoplastic lesions (including hyperplasia and dysplasia) that did not progress to BCCs (Fig. 1b). These results suggest that only IFE-SCs can induce BCC following activation of SmoM2, whereas IFE-CPs are highly resistant to tumour formation.

We then assessed whether the ability of SCs and CPs to initiate BCC was dependent on the oncogene or tumour suppressor gene used to activate HH signalling. To this end, we induced *Ptch1* deletion using K14-CreER or Inv-CreER (Fig. 1c). *Ptch1* deletion using K14-CreER led to BCCs arising from the IFE and the infundibulum (Fig. 1c). In contrast, *Ptch1* deletion using Inv-CreER, which targets some basal cells in the back and ventral skin epidermis[8], did not lead to the rapid development of BCC, and only rare and small BCCs were observed 24 weeks after induction (Fig. 1c, d). These results reveal that only IFE/infundibulum SCs can induce BCC formation, whereas CPs are highly resistant, irrespective of the oncogene or tumour suppressor gene used to activate HH signalling and body location (tail, ear, back and ventral skin).

Two distinct self-maintained compartments, scale and interscale, have been described in tail epidermis[13]. To assess whether cells located in these two compartments respond equally to oncogenic activation, we performed immunofluorescence using a scale-specific marker (K31) and SmoM2–YFP to detect the Smoothened oncogene (SmoM2) fused to YFP on whole-mount tail epidermis. Notably, we found that BCCs arose from K14-CreER SmoM2-targeted cells located only in the interscale (Fig. 1e). K14 clones in the interscale progressively lost their normal differentiation program, as evidenced by the loss of spinous-like cells, became hyperplastic, then dysplastic (Fig. 1f and Extended Data Fig. 1b, c). From 4 to 8 weeks after induction, around 15% of clones had progressed into BCC in interscale, increasing to 40% after 24 weeks (Fig. 1e, f). In contrast, K14 clones in scale never progressed to BCC, and maintained a normal differentiation program for an extended period, despite clonal expansion mediated by SmoM2 expression (Fig. 1e, f and Extended Data Fig. 1b, c).

[1] Université Libre de Bruxelles, IRIBHM, Brussels B-1070, Belgium. [2] Cavendish Laboratory, Department of Physics, J. J. Thomson Avenue, Cambridge CB3 0HE, UK. [3] The Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. [4] Wellcome Trust-Medical Research Council Stem Cell Institute, University of Cambridge, Cambridge CB2 1QR, UK. [5] WELBIO, Université Libre de Bruxelles, Brussels B-1070, Belgium.

*These authors contributed equally to this work.

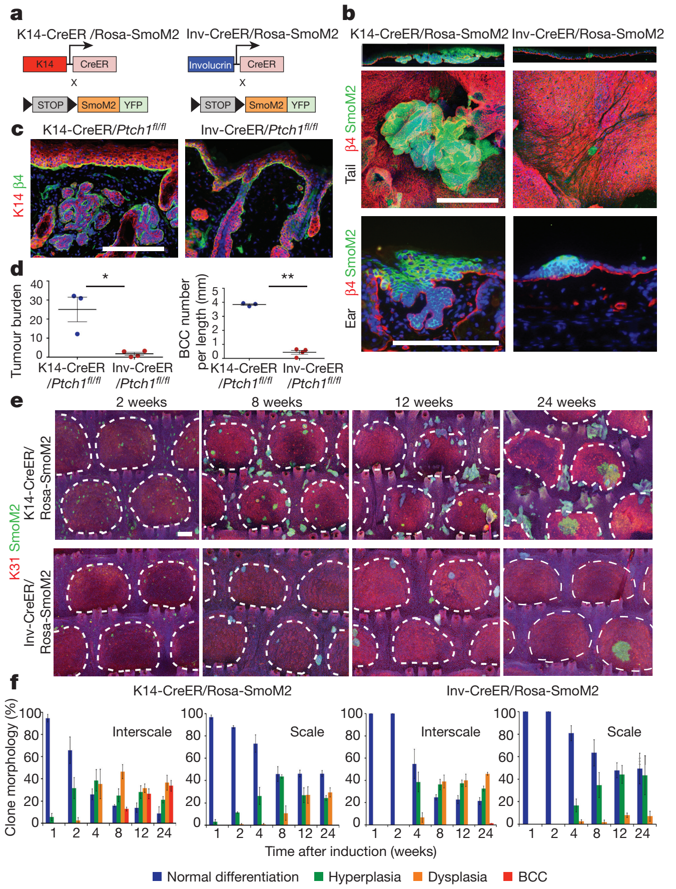


Figure 1 | SCs but not CPs initiate BCC formation upon HH activation. **a**, Genetic strategy to activate SmoM2 expression in SCs and CPs. **b**, Immunostaining of β4-integrin and SmoM2 in ear and tail skin 24 weeks after SmoM2 activation. **c**, Immunostaining of β4-integrin and K14 in ventral skin 24 weeks after *Ptch1* deletion. **d**, Quantification of tumour burden (total tumour area divided by length of epidermis) following *Ptch1* deletion. Quantification of BCC number per length (mm) after *Ptch1* deletion ($n = 4$ Inv-CreER/*Ptch1*$^{fl/fl}$ animals and $n = 3$ K14-CreER/*Ptch1*$^{fl/fl}$ animals). **e**, Immunostaining of K31 and SmoM2 in whole-mount tail skin. **f**, Quantification of the morphology of SmoM2-expressing clones. Description of number of counted clones can be found in the Methods. Hoechst nuclear staining in blue; scale bars, 100 μm. *$P \leq 0.05$, **$P \leq 0.01$. Histograms and error bars represent the mean and the s.e.m.

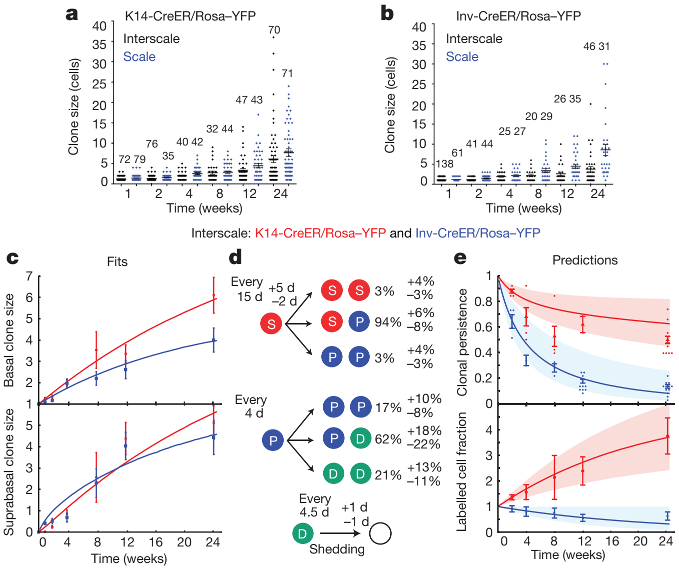


Figure 2 | Homeostatic renewal of mouse tail epidermis. **a**, **b**, Distribution of basal clone sizes, in K14-CreER/Rosa–YFP (**a**) and Inv-CreER/Rosa–YFP (**b**) epidermis. The number of clones analysed is indicated for each time point and described in the Methods. **c**, Mean basal (top) and suprabasal (bottom) clone size in the interscale. The lines represent the model fit. **d**, Cell fate probabilities of SCs and CPs in the interscale, as extracted from the fits. S, P and D refer to stem, progenitor and differentiated cells. **e**, Clonal persistence (top) and labelled cell fraction (bottom) in the interscale. Description of number of counted clones is in the Methods. The lines are the predictions from the model using only the parameters extracted in **d**. K14-CreER/Rosa–YFP clones display a net expansion, whereas Inv-CreER/Rosa–YFP clones display a net contraction. Histograms and error bars represent the mean and the s.e.m. Shaded areas represent 95% confidence intervals for the model prediction (Supplementary Theory).

Together, these data indicate that the fate of oncogene-targeted cells and the ability of these cells to progress into BCC depends both on their location (scale versus interscale) and cellular origin (SC versus CP). This prompted us to investigate whether there are regional differences in SC potential in tail epidermis even under homeostatic conditions.

Homeostasis of the interscale epidermis

To gain quantitative insight into regional variation in SC potential, we performed lineage tracing at homeostasis to determine whether scale and interscale are differentially maintained. To this end, we compared the evolution of K14-CreER/Rosa–YFP-targeted and Inv-CreER/Rosa–YFP-targeted cells at single-cell resolution over a 24-week time course. Interestingly, although both broad, the distributions of clone sizes in the two regions became increasingly divergent (Fig. 2a, b and Extended Data Fig. 2), confirming the importance of regionalization in cellular dynamics (Supplementary Theory).

Consistent with our previous study[7], the evolution of the mean clone size of progenitors targeted by Inv-CreER in the interscale fits well with the targeting of an equipotent CP population presenting a small but statistically significant imbalance in fate towards terminal differentiation (Fig. 2c, d). Similarly, the evolution of mean clone size for K14-CreER cells is consistent with the additional targeting of a long-term self-renewing SC population that divides more slowly than CPs (Fig. 2c, d). To define quantitatively the dynamics of these two populations (cell-cycle times, relative proportion of SCs and CPs labelled by the K14-CreER and their fate probabilities), we made a joint fit to the basal and suprabasal mean clone sizes, and extracted optimal parameters and confidence intervals (Supplementary Theory).

To verify independently the predictions of the model, the persistence of Inv-CreER- and K14-CreER-targeted clones was used to infer the respective labelled cell fractions. As expected from the labelling of the CP population, for Inv-CreER-targeted clones, we found that the labelled cell fraction decreased over time (Fig. 2e). In contrast, for K14-CreER-targeted clones, the labelled cell fraction increased over time, consistent with the preferential targeting of the SC population (Fig. 2e). Notably, we obtained excellent predictions for the labelled cell fraction for both K14-CreER and Inv-CreER using parameters extracted independently from the fit to the mean clone sizes (Fig. 2e). These results support a SC and CP hierarchy, and rule out the possibility that the differences between K14-CreER- and Inv-CreER-targeted clones are the consequence of differential short-term 'priming' of induced cells (Extended Data Fig. 3a). Importantly, the hierarchical model also predicted accurately the complete distribution of clone sizes at all time points (Extended Data Fig. 3b, c) for both K14-CreER and Inv-CreER.

In sharp contrast, in the scale region of tail epidermis, both basal and suprabasal clone sizes and persistence of K14-CreER- and Inv-CreER-targeted cells were statistically indistinguishable (Extended Data Fig. 4a, c). Notably, the labelled cell fraction did not change significantly between 2 weeks and 24 weeks after labelling (Extended Data Fig. 4c), an indication that K14-CreER and Inv-CreER mark the same balanced CP population[13]. We again validated the model (Extended Data Fig. 4b) by showing that it could quantitatively predict both the

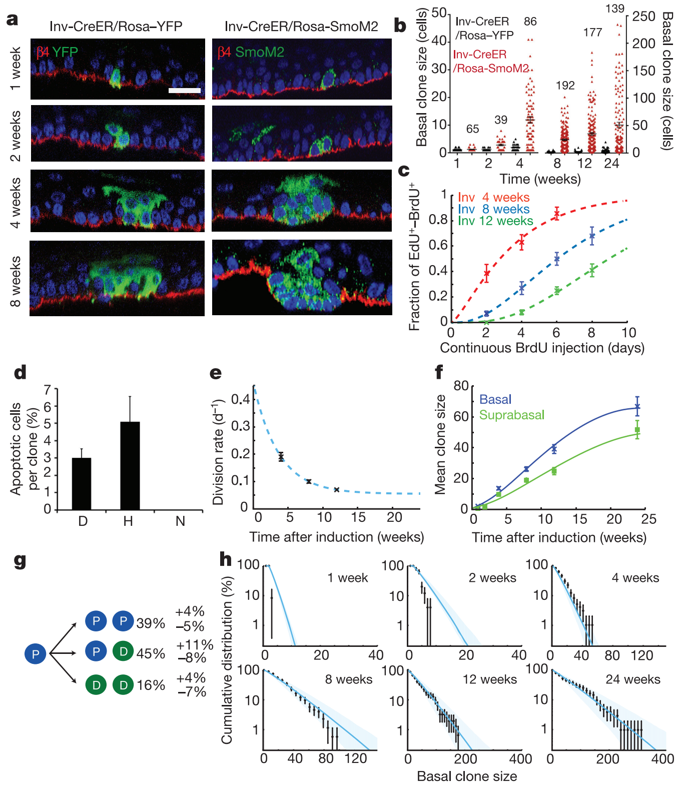


Figure 3 | SmoM2 expression in CPs induces clonal expansion that does not progress into BCC. a, Immunostaining for β4-integrin, YFP and SmoM2 in Inv-CreER/Rosa–YFP and Inv-CreER/Rosa-SmoM2 epidermis at different time points. **b**, Distribution of Inv-CreER/Rosa–YFP and Inv-CreER/Rosa-SmoM2 basal clone sizes. The number of clones analysed for Inv-CreER/Rosa-SmoM2 is indicated for each time point and for Inv-CreER/Rosa–YFP is indicated in Fig. 2b. **c**, Quantification of EdU–BrdU double-labelled cells during continuous BrdU administration, at different time points after clonal induction. The lines represent the model fit (Supplementary Theory). **d**, Quantification of the proportion of apoptotic cells in dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones 8 weeks after induction ($n = 73$ clones analysed from 4 independent experiments). **e**, Division rate (d^{-1}, per day) determined from EdU–BrdU double-labelling experiments (data in black, fit in blue dashed line). **f**, Mean basal and suprabasal clone sizes in the interscale. The lines represent the model fit from which we inferred the cell-fate probabilities displayed in **g**. **g**, Cell-fate probabilities of the tumour progenitor expressing SmoM2. **h**, Basal clone size distribution of Inv-CreER/Rosa-SmoM2 clones (black). Consistent with the hypothesis of a single equipotent progenitor pool, all distributions are well-fit by single exponential. Blue lines represent the model prediction using only the parameters extracted from **g**. Shaded areas represent 95% confidence intervals for the model prediction. D, dysplasia; H, hyperplasia; N, normal differentiation. Hoechst nuclear staining in blue; scale bars, 10 μm. Histograms and error bars represent the mean and the s.e.m. (**b**–**f**) and s.d. (**h**).

evolution of clonal persistence, as well as the clone size distribution at all time points (Extended Data Fig. 4c, d).

These results show that, during homeostasis, interscale is maintained by two discrete populations; a comparatively slow-cycling SC and a more rapidly dividing CP population, whereas scale is maintained by a single CP population. As well as unifying diverging reports of maintenance hierarchy in tail epidermis[7,13,14], these findings raised the question of whether the restriction of BCCs to the interscale correlated with the regional localization of IFE-SCs. To test this hypothesis, we assessed whether the same regionalized lineage hierarchy persisted upon SmoM2 activation.

Oncogene-targeted CPs are frozen into dysplasia

To resolve the cellular dynamics underpinning the differential sensitivity of SCs and CPs to BCC initiation in interscale, we first studied the dynamics and proliferation kinetics of Inv-CreER/Rosa-SmoM2 clones. Oncogenic activation in Inv-CreER CPs lead to an increase of the average basal clone size, total clone size and clonal persistence compared to homeostatic conditions (Fig. 3a, b and Extended Data Fig. 5a–c), as well as abnormal or decreased differentiation (Fig. 3a and Extended Data Fig. 1 b, c). We assessed the average cell-cycle time of SmoM2 Inv-CreER-targeted cells by first marking proliferating cells using 24 h of EdU administration, followed by variable periods of continuous BrdU administration. From the co-labelling of EdU–BrdU, we found that CPs divided on average every 3.6 ± 0.5 days 4 weeks after SmoM2 expression, 7.2 ± 0.6 days after 8 weeks and 9.8 ± 0.3 days after 12 weeks (Fig. 3c), indicating that the average division rate of SmoM2 CPs decreases with time. Surprisingly, division rates were uncorrelated with clone size at all time points, indicating that the decrease occurs independently of clone size or stage of tumour progression (Extended Data Fig. 5d), and consistent with the Inv-CreER oncogene-targeted cells functioning as a single equipotent population.

As deregulation of apoptosis is also important for cancer formation[1], we assessed whether apoptosis influences the clonal dynamics of oncogene-targeted CPs. In common with their normal counterpart, Inv-CreER-targeted cells did not show evidence of apoptosis over the first 6 weeks after SmoM2 expression (data not shown). However, from 8 weeks on, about 60% of Inv-CreER-targeted clones that presented hyperplasia or dysplasia, contained about 2–4% of apoptotic cells as measured by active caspase-3 immunostaining (Fig. 3d and Extended Data Fig. 5e–i).

Taking these rates (Fig. 3e) as an input, we could obtain an excellent fit to the average clone size (Fig. 3f) with cell fate probabilities that remain constant over time, with the proportions of symmetric renewal (PP) to asymmetric division (PD) and symmetric differentiation (DD) set at 39%, 45% and 16% respectively (Fig. 3g). This result demonstrates that oncogenic expression in CPs leads to enhanced clonal expansion and survival by promoting symmetric proliferation over terminal differentiation. Such an imbalance would lead to exponential clone growth if it were not counteracted by an ever-diminishing effective proliferation rate, leading to a plateau in the mean basal clone size (Fig. 3f). Notably, the model prediction provided a good fit to the clone size distribution at all time points (Fig. 3h).

Finally, to further verify the model, a low short-term dose of EdU was used to mark a minority of dividing cells and their fate outcome was recorded 3 days later by quantifying the basal and suprabasal localization of EdU doublets (Extended Data Fig. 5j). From these results, we could confirm a large imbalance between symmetric division and terminal differentiation (35%).

As the scale is maintained by a single progenitor pool, we investigated whether its response to oncogenic activation was similar to interscale CPs. Notably, after an initial increase, the overall labelled cell fraction remained roughly constant over time in scale between 8 and 24 weeks, at a similar level for both K14-CreER and Inv-CreER (Extended Data Fig. 6a–c), suggesting that, in sharp contrast to interscale, both populations behave identically upon oncogenic activation (Extended Data Fig. 6d). Together, these results show that both interscale and scale CPs are resistant to BCC formation upon oncogenic HH signalling, although interscale clones can persist longer owing to a larger fate imbalance and enhanced differentiation defects, whereas scale clones rapidly converge towards balance. However, as human epidermis does not show scale organization, the absence of BCC formation in the scale region might not have human relevance.

Oncogene-targeted SCs progress into BCC

To gain insight into how SmoM2 expression in SCs promotes BCC formation, we then performed a quantitative analysis of K14-CreER/Rosa-SmoM2 clones. Compared to Inv-targeted clones, SmoM2 expression in K14-targeted cells lead to a more rapid and persistent expansion of a fraction of clones (Fig. 4a, b and Extended Data Fig. 7a–c) that progressed into BCC, as well as the formation of smaller clones that did not

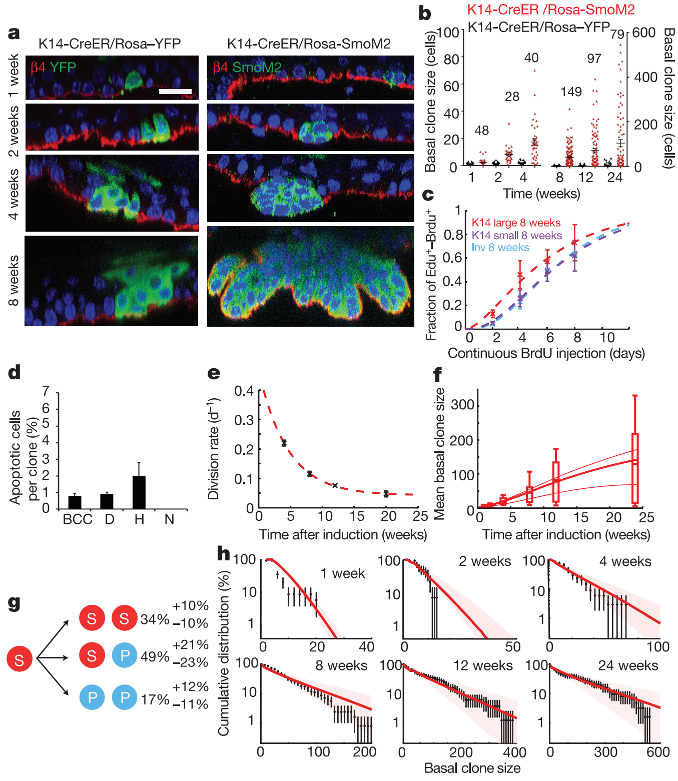


Figure 4 | SmoM2 expression in SCs induces tumour SCs that lead to BCC formation. a, Immunostaining for β4-integrin, YFP and SmoM2 in K14-CreER/Rosa–YFP and K14-CreER/Rosa-SmoM2 epidermis at different time points. **b**, Distribution of K14-CreER/Rosa–YFP and K14-CreER/Rosa-SmoM2 basal clone sizes. The number of clones analysed for K14-CreER/Rosa-SmoM2 is indicated and for K14-CreER/Rosa–YFP is indicated in Fig. 2a. **c**, Quantification of EdU–BrdU double-labelled cells following continuous BrdU administration, at 8 weeks after clonal induction for small K14-CreER clones, Inv-CreER clones, and large K14-CreER clones. **d**, Quantification of the number of apoptotic cells in BCC, dysplastic, hyperplastic and normally differentiating K14-CreER/Rosa-SmoM2 clones 8 weeks after induction ($n = 117$ clones analysed from 4 independent experiments). **e**, Division rate (d^{-1}, per day) in large K14 clones determined from double-labelling experiments (data in black, fit in red dashed line). **f**, Whisker plot of the mean basal clone size in the interscale. The boxes delineate the first and third quartiles of the data and the whiskers delineate the first and last deciles of the data. The thick continuous line is the best fit from the model from which we extract the probability of fate choices in tumour SCs and progenitors displayed in **g**. The thin lines represent the predicted mean clone sizes of SC- (top thin curve) and CP- (bottom thin curve) derived clones alone. **g**, Cell-fate probabilities of the tumour SC upon SmoM2 activation. **h**, Basal clone size distribution of K14-CreER/SmoM2 clones (black). Red lines are the model prediction using only the parameters extracted from **g**. Shaded areas represent 95% confidence intervals for the model prediction. Hoechst nuclear staining in blue; scale bars, 10 μm. Histograms and error bars represent the mean and the s.e.m. (**b**–**f**) and the s.d. (**h**).

show tumour progression (Figs 1e, 4b). This suggests that, in line with homeostatic conditions, K14-CreER marks a fraction of tumour-like SCs, together with tumour-like CPs, a heterogeneity that we verified using proliferation assays (Fig. 4c). Indeed, we found that one population of K14-CreER-targeted cells consisted of small clones that displayed similar proliferation kinetics as Inv-CreER SmoM2 clones, whereas a second population consisted of larger clones, which re-entered cell cycle significantly faster (Fig. 4c and Extended Data Figs 5d and 7d). The population of small K14-CreER-targeted clones (hyperplasia) also presented higher levels of apoptosis compared to the larger clones (Fig. 4d and Extended Data Fig. 7e–i). As a result, even though the proliferation of the larger clones also decreased with time (Fig. 4e and Extended Data Fig. 7j), their division rate was consistently higher than the Inv-CreER-targeted population (Fig. 4e).

To model BCC initiation, we adapted the hierarchical model obtained during homeostasis and fitted jointly the mean basal and suprabasal clone sizes of all K14-CreER SmoM2 clones, taking as input the division rate as well as the fraction of SCs initially labelled by the K14-CreER-determined from measurements at homeostasis, and used the fate choices of SCs as fitting parameters (Fig. 4f and Extended Data Fig. 7k). In particular, we posited that SCs are imbalanced towards symmetric renewal, whereas CPs derived from these cells remain slightly imbalanced towards symmetric differentiation with the same fate probability as in homeostasis, which gave a good fit to the average basal clone size (Fig. 4g). Notably, the measured clone size distributions from 12 weeks onwards could not be fit with a one-progenitor population model, in contrast to the distributions of Inv-CreER SmoM2 clones. Instead, the K14-CreER SmoM2 clone size distributions displayed a 'double-exponential' decay, consistent with the labelling of two distinct populations, as predicted quantitatively by the model (Fig. 4h and Supplementary Theory). This shows that K14-CreER targets tumour-like SCs making imbalanced stochastic fate choices, in addition to targeting the same tumour-like CP population as Inv-CreER.

As a final consistency check, we addressed a key hallmark of the hierarchical model, that SCs give rise to basal CPs in K14-CreER-targeted clones. This predicts that the fraction of cell divisions resulting in two basal cells should be greater in SC- versus CP-targeted clones. Indeed, short-term EdU pulse-chase experiments revealed that, in BCC, most divisions (77%) lead to two basal cells (Extended Data Fig. 7l). In hyperplasia/dysplasia, the fraction of two EdU^{+} basal cell doublets was intermediate between the BCC and Inv-CreER/Rosa-SmoM2 values (Extended Data Figs 5j, 7l), consistent with a mixture of SC- and CP-targeted clones.

p53 restricts CP progression to BCC

Given the observed differences in apoptosis and division rates between oncogene-targeted SCs and CPs, we assessed whether *p53*, a tumour suppressor gene frequently mutated in human BCC[15] that controls cell cycle arrest and apoptosis[16], was differentially activated in SCs and CPs upon SmoM2 activation. Immunohistochemistry revealed that p53 was more frequently found in SmoM2 clones arising from Inv-CreER as compared to K14-CreER mice (Extended data Fig. 8a). To determine whether p53 stabilization in oncogene-targeted CPs restricts the potential of these progenitors to generate BCC, we deleted *p53* together with SmoM2 activation and assessed tumour formation. Interestingly, *p53* deletion in Inv-CreER targeted CPs leads to BCC in both ear and tail epidermis (Fig. 5a). In the tail, BCCs were restricted to the interscale whereas, in the scale, clones only progressed into dysplasia (Fig. 5b, c and Extended Data Fig. 8b). These results indicate that p53 restricts the competence of SmoM2-targeted CPs of the interscale to progress into BCC.

Although the proportion of clones that progress into BCC continued to be more frequent and more rapid in K14-CreER targeted SCs, at 24 weeks after induction more than half of interscale Inv-CreER-targeted clones had progressed into BCC after *p53* deletion (Fig. 5c). The clonal persistence and clone size were increased upon *p53* deletion in both Inv-CreER/Rosa-SmoM2 and K14-CreER/Rosa-SmoM2 interscale clones, although the clones were still bigger and more persistent in K14-targeted cells (Fig. 5b, d and Extended Data Fig. 8c–e). These results indicate that, upon *p53* deletion, both oncogene-targeted CPs and SCs present an increase in self-renewing divisions allowing CPs to acquire the ability to form BCC upon SmoM2 expression.

We next determined whether the observed increase in clone size in the absence of p53 in CPs and SCs was due to a decrease in apoptosis, an increase in proliferation or both. Immunostaining for active caspase-3 8 weeks after oncogenic activation showed that large Inv-CreER/Rosa-SmoM2/$p53^{fl/fl}$ dysplastic and BCC clones displayed reduced apoptosis,

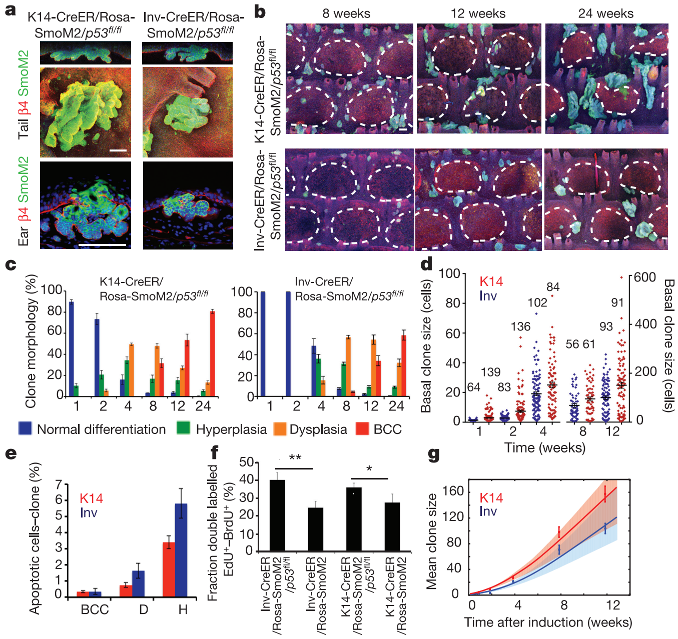


Figure 5 | p53 deletion in CPs leads to BCC formation. a, Immunostaining of β4-integrin and SmoM2 in ear and tail skin of K14 and Inv-CreER/Rosa-SmoM2/$p53^{fl/fl}$ mice 24 weeks after tamoxifen administration. **b**, Whole-mount immunostaining of K31/SmoM2 in tail epidermis over time. **c**, Quantification of normal, hyperplastic, dysplastic and BCC clones in the interscale region. Description of number of counted clones is found in the Methods section. **d**, Distribution of basal clone sizes in K14 and Inv-CreER/Rosa-SmoM2/$p53^{fl/fl}$ mice. The number of clones analysed is indicated. Clone merger events were observed 12 weeks after oncogenic activation in K14Cre-ER/Rosa-SmoM2/$p53^{fl/fl}$ preventing the accurate quantification of clonal persistence and clone size at longer times. **e**, Quantification of the proportion of apoptotic cells in different clones (K14 $n = 82$ clones and Inv $n = 90$ clones from 3 independent experiments). **f**, Percentage of double-labelled EdU–BrdU SmoM2-expressing cells after 6 days of continuous BrdU administration following a 24 h pulse of EdU at 12 weeks after induction. $*P \leq 0.05$, $**P \leq 0.01$. **g**, Mean basal clone size in Inv-CreER/Rosa-SmoM2/$p53^{fl/fl}$ and K14-CreER/Rosa-SmoM2/$p53^{fl/fl}$ clones. The prediction of the model is indicated by the blue and red lines. Shaded areas represent 95% confidence intervals for the model prediction in **g**. Hoechst nuclear staining in blue; scale bars, 100 μm. Histograms and error bars represent the mean and the s.e.m.

mirroring our observation in K14-CreER/Rosa-SmoM2 (Extended Data Fig. 8f). However, apoptosis was unchanged in Inv-CreER p53-deficient hyperplastic clones, suggesting that p53-dependent and -independent mechanisms control apoptosis in oncogene-targeted cells (Fig. 5e, Extended Data Fig. 8g). EdU–BrdU double-pulse experiments 12 weeks after induction showed that deletion of *p53* increased the rate of proliferation in both Inv-CreER and K14-CreER oncogene-targeted cells (Fig. 5f). According to our model, this increase in the rate of division was sufficient, keeping all other parameters constant, to explain the enhanced tumour growth (Fig. 5g). This provides additional evidence that growth arrest in oncogene-targeted CPs is a key determinant in their inability to mediate BCC progression in the presence of p53.

In summary, our results demonstrate that p53 restricts the ability of CPs to initiate BCC by promoting apoptosis and inducing cell cycle arrest in oncogene-targeted CPs.

Discussion

In this study, we have defined the quantitative dynamics of BCC initiation at single-cell resolution, from the first oncogenic hit to the development of invasive tumours. These results show that the proliferative hierarchical organization of skin epidermis is a key determinant of tumour development, with only IFE-SCs and not CPs able to initiate BCC following oncogenic HH signalling (Extended data Fig. 9). Even though CP-derived clones survive and proliferate for months, they are surprisingly robust to BCC transformation and invasion, becoming ‘frozen’ in a pre-tumorigenic state. The developmental cerebellar progenitors initiate medulloblastoma upon oncogenic HH signalling[17,18], suggesting the developmental stage of progenitors may also dictate competence for tumour initiation. The long-term maintenance of some oncogene-targeted CPs contrasts the classical transient-amplifying cells in other compartments, such as hair matrix in the skin or the non-Lgr5 crypt progenitors in gut, which are resistant to tumour initiation because of their short lifespan[19–22].

Our results show that IFE-SCs reside solely in the interscale region, and have the unique and regionalized competence to initiate large and invasive BCCs. Notably, this regionalized hierarchical organization at homeostasis was maintained upon SmoM2 activation. Oncogene expression in SCs lead to a more rapid clonal expansion as compared to CPs for two main reasons: the maintenance of hierarchical organization in early pre-neoplastic lesions, leading to increased symmetric self-renewing divisions; and the combined resistance to apoptosis and enhanced proliferation of SC-derived pre-neoplastic lesions, leading to a larger effective growth rate. These two properties allow SC-targeted tumours to escape the frozen state that characterized CP-targeted pre-neoplastic lesions, and thereby progress to an invasive phenotype.

Finally, our results show that p53 restricts the ability of CPs to undergo BCC initiation by promoting apoptosis and inducing cell-cycle arrest in oncogene-targeted CPs. Interestingly, although the division rates of CPs and SCs that are p53-deficient are similar, SC-targeted tumours still grow to larger sizes than CP-targeted tumours, suggesting that the hierarchical organization is at least partially maintained even after two oncogenic hits. By establishing that sustained imbalance towards self-renewing divisions and resistance to p53-mediated apoptosis and cell-cycle arrest are the main drivers of tumorigenesis, our results suggest that therapy promoting differentiation, p53 reactivation and apoptosis could present a promising avenue to promote BCC regression and prevent tumour relapse.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 January; accepted 4 July 2016.
Published online 8 July 2016.

1. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: the next generation. *Cell* **144,** 646–674 (2011).
2. Vermeulen, L. *et al.* Defining stem cell dynamics in models of intestinal tumor initiation. *Science* **342,** 995–998 (2013).
3. Snippert, H. J., Schepers, A. G., van Es, J. H., Simons, B. D. & Clevers, H. Biased competition between Lgr5 intestinal stem cells driven by oncogenic mutation induces clonal expansion. *EMBO Rep.* **15,** 62–69 (2014).
4. Epstein, E. H. Basal cell carcinomas: attack of the hedgehog. *Nat. Rev. Cancer* **8,** 743–754 (2008).
5. Blanpain, C. & Simons, B. D. Unravelling stem cell dynamics by lineage tracing. *Nat. Rev. Mol. Cell Biol.* **14,** 489–502 (2013).
6. Blanpain, C. & Fuchs, E. Stem cell plasticity. Plasticity of epithelial stem cells in tissue regeneration. *Science* **344,** 1242281 (2014).
7. Mascré, G. *et al.* Distinct contribution of stem and progenitor cells to epidermal maintenance. *Nature* **489,** 257–262 (2012).
8. Lapouge, G. *et al.* Identifying the cellular origin of squamous skin tumors. *Proc. Natl Acad. Sci. USA* **108,** 7431–7436 (2011).
9. Youssef, K. K. *et al.* Identification of the cell lineage at the origin of basal cell carcinoma. *Nat. Cell Biol.* **12,** 299–305 (2010).
10. Youssef, K. K. *et al.* Adult interfollicular tumour-initiating cells are reprogrammed into an embryonic hair follicle progenitor-like fate during basal cell carcinoma initiation. *Nat. Cell Biol.* **14,** 1282–1294 (2012).
11. Wong, S. Y. & Reiter, J. F. Wounding mobilizes hair follicle stem cells to form tumors. *Proc. Natl Acad. Sci. USA* **108,** 4093–4098 (2011).
12. Kasper, M. *et al.* Wounding enhances epidermal tumorigenesis by recruiting hair follicle keratinocytes. *Proc. Natl Acad. Sci. USA* **108,** 4099–4104 (2011).
13. Gomez, C. *et al.* The interfollicular epidermis of adult mouse tail comprises two distinct cell lineages that are differentially regulated by Wnt, Edaradd, and Lrig1. *Stem Cell Reports* **1,** 19–27 (2013).
14. Clayton, E. *et al.* A single type of progenitor cell maintains normal epidermis. *Nature* **446,** 185–189 (2007).

15. Bonilla, X. *et al.* Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma. *Nat. Genet.* **48,** 398–406 (2016).
16. Chen, J. The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression. *Cold Spring Harb. Perspect. Med.* **6,** a026104 (2016).
17. Schüller, U. *et al.* Acquisition of granule neuron precursor identity is a critical determinant of progenitor cell competence to form Shh-induced medulloblastoma. *Cancer Cell* **14,** 123–134 (2008).
18. Yang, Z. J. *et al.* Medulloblastoma can be initiated by deletion of Patched in lineage-restricted progenitors or stem cells. *Cancer Cell* **14,** 135–145 (2008).
19. Barker, N. *et al.* Crypt stem cells as the cells-of-origin of intestinal cancer. *Nature* **457,** 608–611 (2009).
20. White, A. C. *et al.* Defining the origins of Ras/p53-mediated squamous cell carcinoma. *Proc. Natl Acad. Sci. USA* **108,** 7425–7430 (2011).
21. Lapouge, G. *et al.* Skin squamous cell carcinoma propagating cells increase with tumour progression and invasiveness. *EMBO J.* **31,** 4563–4575 (2012).
22. Zhu, L. *et al.* Prominin 1 marks intestinal stem cells that are susceptible to neoplastic transformation. *Nature* **457,** 603–607 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements We would like to thank J.-M. Vanderwinden and the LiMiF for the help with confocal microscopy. C.B. is an investigator of WELBIO. A.S.-D. and J.C.L. are supported by a fellowship of the FNRS and FRIA respectively. B.D.S. and E.H. are supported by the Wellcome Trust (grant numbers 098357/Z/12/Z and 110326/Z/15/Z). E.H. is supported by a fellowship from Trinity College, Cambridge. This work was supported by the FNRS, the IUAP program, the Fondation contre le Cancer, the ULB fondation, the foundation Bettencourt Schueller, the foundation Baillet Latour, a consolidator grant of the European Research Council.

Author Contributions A.S.-D., C.B., E.H. and B.D.S. designed the experiments, performed data analysis and wrote the manuscript; A.S.-D. performed all of the biological experiments; E.H. performed all the mathematical modelling. J.C.L. and M.L. provided technical support. K.K.Y. made initial observations pertinent to the study.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.B. (Cedric.Blanpain@ulb.ac.be) and B.D.S (bds10@cam.ac.uk).

Reviewer Information *Nature* thanks S. Benitah, F. de Sauvage, L. Vermeulen and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Mice. K14-CreER transgenic mice[23] were kindly provided by E. Fuchs, Rockefeller University; Inv-CreER were generated in our laboratory[8]. *Ptch1*$^{fl/fl}$ mice[24] and Rosa/SmoM2-YFP mice[25] were obtained from the JAX repository. *p53*$^{fl/fl}$ (ref. 26) mice were obtained from the National Cancer Institute at Frederick.

Mouse colonies were maintained in a certified animal facility in accordance with European guidelines. Experiments involving mice presented in this work were approved by Comité d'Ethique du Bien Être Animal (Université Libre de Bruxelles) under protocol number 483N, that states that animals should be euthanized if they present tumours that exceed 1 cm in diameter. The BCCs observed in this study were microscopic and ranged from 1.5 mm to 100 μm in diameter and in none of the experiments performed, the tumours exceeded the limit (1 cm in diameter) described in protocol 483N. Female and male animals have been used for all experiments and equal animal gender ratios have been respected in the majority of the analysis, analysis of the different mutant mice was not blind and sample size was calculated to reach statistical significance. The experiments were not randomized.

Skin tumour induction and clonal YFP expression. For clonal induction 3-months-old mice were used. K14-CreER/Rosa-YFP, K14-CreER/Rosa-SmoM2, K14-CreER/SmoM2/*p53*$^{fl/fl}$ and K14-CreER/*Ptch1*$^{fl/fl}$ mice received an intraperitoneal injection of 0.1 mg of tamoxifen and Inv-CreER/Rosa-YFP, Inv-CreER/Rosa-SmoM2, Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ and Inv-CreER/*Ptch1*$^{fl/fl}$ received a intraperitoneal injection of 2.5 mg of tamoxifen to achieve similar level of recombination in the different models (Extended Data Fig. 1a). Mice were killed and analysed at different time points following tamoxifen administration.

Immunostaining in sections. The tail, ventral skin and ear skin were embedded in optimal cutting temperature compound (OCT, Sakura) and cut into 5–8 μm frozen sections using a CM3050S Leica cryostat (Leica Microsystems).

Immunostainings were performed on frozen sections. Owing to the fusion of SmoM2 with YFP, SmoM2-expressing cells were detected using anti-GFP antibody. Frozen sections were dried and then fixed with 4% paraformaldehyde/PBS (PFA) for 10 min at room temperature and blocked with blocking buffer for 1 h (PBS, horse serum 5%, BSA 1%, Triton 0.1%). Skin sections were incubated with primary antibodies diluted in blocking buffer overnight at 4 °C, washed with PBS for 3 × 5 min, and then incubated with Hoechst solution and secondary antibodies diluted in blocking buffer for 1 h at room temperature. Finally, sections were washed with PBS for 3 × 5 min at room temperature and mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma). Primary antibodies used were the following: anti-GFP (rabbit, 1:1,000, BD, A11122), anti-K14 (Chicken, 1:4,000, Covance, PCK-153P-0100) and anti-B4-integrin (rat, 1:200, BD, 553745). The following secondary antibodies were used: anti-rabbit, anti-rat, anti-chicken, conjugated to AlexaFluor488 (Molecular Probes) and to rhodamine Red-X (JacksonImmunoResearch). Images of immunostaining in sections were acquired using an Axio Imager M1 microscope, an AxioCamMR3 camera and the Axiovision software (Carl Zeiss).

Immunostaining in whole mounts. Whole mounts of tail epidermis were performed as previously described[27] and used to quantify the proportion of surviving clones (Extended Data Fig. 2b) as well as the basal suprabasal and total clone size. Specifically, pieces of tail were incubated for 1 h at 37 °C in EDTA 20 mM in PBS in a rocking plate, then using forceps the dermis and epidermis were separated and the epidermis was fixed for 30 min in PFA 4% in agitation at room temperature and washed 3 times with PBS.

For the immunostaining, tail skin pieces were blocked with blocking buffer for 3 h (PBS, horse serum 5%, Triton 0.8%) in a rocking plate at room temperature. After, the skin pieces were incubated with primary antibodies diluted in blocking buffer overnight at 4 °C, the next day they were washed with PBS-Tween 0.2% for 3 × 10 min at room temperature, and then incubated with the secondary antibodies diluted in blocking buffer for 3 h at room temperature, washed 2 × 10 min with PBS-Tween 0.2% and washed for 10 min in PBS. Finally, they were incubated in Hoechst diluted in PBS for 30 min at room temperature in the rocking plate, washed 3 × 10 min in PBS and mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma). Primary antibodies used were the following: anti-GFP (rabbit, 1:100, BD, A11122), anti-GFP (goat, 1:800, Abcam, Ab6673), anti-active-caspase3 (rabbit, 1:600, R&D, AF835), anti-β4-integrin (rat, 1:200, BD, 553745) and anti-K31 (guinea pig, 1:200, Progen, GP-hHa1). The following secondary antibodies were used: anti-rabbit, anti-rat, anti-chicken, anti-goat and anti-guinea pig, conjugated to AlexaFluor488 (Molecular Probes), to rhodamine Red-X (JacksonImmunoResearch) and to Cy5 (1:400, Jackson ImmunoResearch).

Analysis of clone survival, size and apoptosis. Quantification of the proportion of surviving clones, as well as total and basal clone size was determined by counting the number of SmoM2–YFP and YFP-positive cells. in each clone using whole-mount tail epidermis. The different clones were imaged using *Z*-stacks using a confocal microscope LSM 780 (Carl Zeiss) and orthogonal views were used to count the number of basal and total number of SmoM2–YFP or YFP-positive cells in each clone, as well as the number of active-caspase3-positive cells in each clone. K31 staining was used to classify the clones according to their location in the scale or interscale regions.

Proliferation assays. To measure the kinetics of cell proliferation, a 24 h continuous pulse of EdU followed with a continuous pulse of BrdU were performed. Specifically, mice received at $t = 0$ an intraperitoneal injection of EdU (1 $\mathrm{mg\,ml^{-1}}$) and 0.1 $\mathrm{mg\,ml^{-1}}$ EdU was added to their drinking water for 24 h. The next days the mice received a daily intraperitoneal injection of BrdU (10 $\mathrm{mg\,ml^{-1}}$) and 1 $\mathrm{mg\,ml^{-1}}$ of BrdU was added to their drinking water during the 8 days of the continuous BrdU pulse. Mice were killed at different time points and whole-mount stainings for the tail were performed. The pieces of tail were first stained for GFP (following the protocol described in the previous section). Second, EdU staining was performed following the manufacturer's instructions (Invitrogen). The pieces of tail were then washed in PBS and fixed again in PFA 4% for 10 min. After they were washed in PBS, incubated for 20 min in HCl 1 M at 37 °C, washed three times with PBS-Tween 0.2% and incubated overnight with Alexa-647-coupled anti-BrdU antibody (mouse, 1:200, BD). The next day the tail pieces were washed in PBS, incubated in Hoechst for 30 min at room temperature in the rocking plate, washed 3 × 10 min in PBS and mounted in DAKO mounting medium supplemented with 2.5% Dabco (Sigma). To quantify the number of cells that incorporated EdU and/or BrdU, *Z*-stacks were acquired for each individual clone and orthogonal views used to count.

Immunohistochemistry. For p53 immunohistochemistry, 4-μm paraffin sections were deparaffinized, rehydrated, followed by antigen unmasking performed for 20 min at 98 °C in citrate buffer (pH 6) using the PT module. Endogenous peroxydase was blocked using 3% H_2O_2 (Merck) in methanol for 10 min at room temperature. Endogenous avidin and biotin were blocked using the Endogenous Blocking kit (Invitrogen) for 20 min at room temperature. In p53 staining, nonspecific antigen blocking was performed using M.O.M. Basic kit reagent. Mouse anti-p53 antibody (clone 1C12; Cell Signaling) was incubated overnight at 4 °C. Anti-mouse biotinylated with M.O.M. Blocking kit, Standard ABC kit, and ImmPACT DAB (Vector Laboratories) was used for the detection of horseradish peroxidase (HRP) activity. Slides were then dehydrated and mounted using SafeMount (Labonord).

Supplementary statistics. For the quantification of the clone morphology of SmoM2-expressing clones in the scale and interscale regions (Fig. 1f), we counted in K14-CreER/Rosa-SmoM2 mice, 128, 109, 76, 195, 168 and 142 clones in the interscale region; 141, 116, 74, 94, 78 and 69 clones in the scale region from 3, 4, 4, 6, 4 and 5 independent experiments at 1, 2, 4, 8, 12 and 24 w respectively. In Inv-CreER/Rosa-SmoM2 mice, 104, 78, 42, 127, 160 and 344 clones were counted in the interscale region; 94, 54, 99, 90, 99 and 39 clones in the scale region from 4, 4, 4, 5, 4 and 8 independent experiments at 1, 2, 4, 8, 12 and 24 weeks, respectively.

For the analysis of the clone size of the K14-CreER/Rosa-YFP mice (Fig. 2a, c and Extended Data Fig. 2), we counted clones (both in scale and interscale) from two independent experiments at 1 week and 2 weeks, five independent experiments at 4 weeks, three independent experiments at 8 weeks, two independent experiments at 12 weeks and four independent experiments at 24 weeks. For the analysis of the clone size of the Inv-CreER/Rosa-YFP mice (Fig. 2b, c, Extended Data Fig. 2), we counted clones (both in scale and interscale) from two independent experiments at 1 week and 2 weeks, five independent experiments at 4 weeks, three independent experiments at 8 weeks, four independent experiments at 12 weeks and three independent experiments at 24 weeks (see raw data in cited figures (Source Data)).

For the clonal persistence of the K14-CreER/Rosa-YFP mice (Fig. 2e and Extended Data Fig. 2), we counted 167, 176, 129, 100, 47 and 246 clones in interscale and 184, 109, 75, 66, 19 and 103 clones in scale from 4, 5, 5, 5, 2 and 4 independent experiments at 1, 2, 4, 8, 12 and 24 w respectively. For 24 weeks, we counted several areas per mice as the number of clones was reduced (see Source Data).

For the clonal persistence of the Inv-CreER/Rosa-YFP mice (Fig. 2e and Extended Data Fig. 2), we counted 138, 95, 25, 31, 76 and 54 clones in interscale and 12, 17, 7, 8, 20 and 10 clones in scale from 2, 4, 2, 3, 4 and 3 independent experiments at 1, 2, 4, 8, 12 and 24 weeks respectively. For 12 and 24 weeks, we counted several areas per mice as the number of clones was low (see Source Data).

For the analysis of the clone size of the Inv-CreER/Rosa-SmoM2 mice (Fig. 3b, f, h, Extended Data Figs 5, 6), we counted clones (both in scale and interscale) from two independent experiments at 1 week and 2 weeks, from four independent experiments at 4 weeks, from six independent experiments at 8 weeks, from six independent experiments at 12 weeks and from four independent experiments at 24 weeks (see Source Data).

For the clonal persistence of the Inv-CreER/Rosa-SmoM2 mice (Extended Data Figs 5, 6), we counted 65, 39, 71, 51, 27, 18 clones in interscale and 67, 27, 47, 31, 12 and 6 clones in scale from 2, 2, 4, 3, 2 and 2 independent experiments at 1, 2, 4, 8, 12 and 24 weeks, respectively (see Source Data).

For the analysis of the clone size of the K14-CreER/Rosa-SmoM2 mice (Fig. 4b, f, h and Extended Data Figs 6, 7), we counted clones (both in scale and interscale) from three independent experiments at 1 week, from two independent experiments at 2 weeks, 4 weeks, from six independent experiments at 8 weeks, from four independent experiments at 12 weeks and from two independent experiments at 24 weeks (see Source Data).

For the clonal persistence of the K14-CreER/Rosa-SmoM2 mice (Extended Data Figs 6, 7), we counted 122, 63, 81, 79, 74 and 68 clones in interscale and 89, 46, 37, 42, 31 and 16 clones in scale from 4, 3, 4, 4, 4 and 4 independent experiments at 1, 2, 4, 8, 12 and 24 weeks respectively (see Source Data).

For the cell proliferation kinetics experiments in the Inv-CreER/Rosa-SmoM2 mice (Fig. 3c, e): at 4 weeks after induction, we counted 33 clones from 3 independent experiments for 2 days of continuous BrdU, 30 clones from 2 independent experiments for 4 days of continuous BrdU, 33 clones from 2 independent experiments for 6 days of continuous BrdU. At 8 weeks after induction, we counted 41 clones from $n=3$ mice for 2 days of continuous BrdU, 16 clones from 2 independent experiments for 4 days of continuous BrdU, 30 clones from 2 independent experiments for 6 days of continuous BrdU and 24 clones from 2 independent experiments for 8 days of continuous BrdU. At 12 weeks after induction, we counted 19 clones from 2 independent experiments for 2 days of continuous BrdU, 26 clones from 2 independent experiments for 4 days of continuous BrdU, 27 clones from 2 independent experiments for 6 days of continuous BrdU and 31 clones from 2 independent experiments mice for 8 days of continuous BrdU. For the 2 weeks after induction data point, we use solely continuous BrdU incorporation, and counted 54 clones from two independent experiments.

For the cell proliferation kinetics experiments in the K14-CreER/Rosa-SmoM2 mice (Fig. 4c, e): at 4 weeks after induction, we counted 56 clones from 3 independent experiments for 2 days of continuous BrdU, 39 clones from 3 independent experiments for 4 days of continuous BrdU, 29 clones from 3 independent experiments for 6 days of continuous BrdU. At 8 weeks after induction, we counted 30 clones from 2 independent experiments for 2 days of continuous BrdU, 25 clones from 2 independent experiments for 4 days of continuous BrdU, 63 clones from 3 independent experiments for 6 days of continuous BrdU and 41 clones from 3 independent experiments for 8 days of continuous BrdU. At 12 weeks after induction, we counted 20 clones from 2 independent experiments for 2 days of continuous BrdU, 21 clones from 2 independent experiments for 4 days of continuous BrdU, 28 clones from 2 independent experiments for 6 days of continuous BrdU and 26 clones from two independent experiments for 8 days of continuous BrdU.

For the quantification of the clone morphology in absence of p53 interscale (Fig. 5c). For K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice 186, 217, 90, 343, 452 and 543 clones from 3, 3, 2, 3, 5 and 5 independent experiments and for Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ 95, 98, 199, 271, 263 and 210 clones from 3, 3, 3, 4, 4 and 4 independent experiments were analysed at 1, 2, 4, 8, 12 and 24 weeks respectively. In the quantification in the scale region (Extended Data Fig. 8b) for K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ 178, 204, 100, 132, 232 and 120 clones were counted from 3, 3, 2, 3, 5 and 5 independent experiments 1, 2, 4, 8, 12 and 24 weeks respectively. For Inv-CreER/Rosa-SmoM2 82, 127, 167, 136, 62 and 153 clones were counted from 2, 3, 3, 4, 4 and 5 independent experiments 1, 2, 4, 8, 12 and 24 weeks respectively.

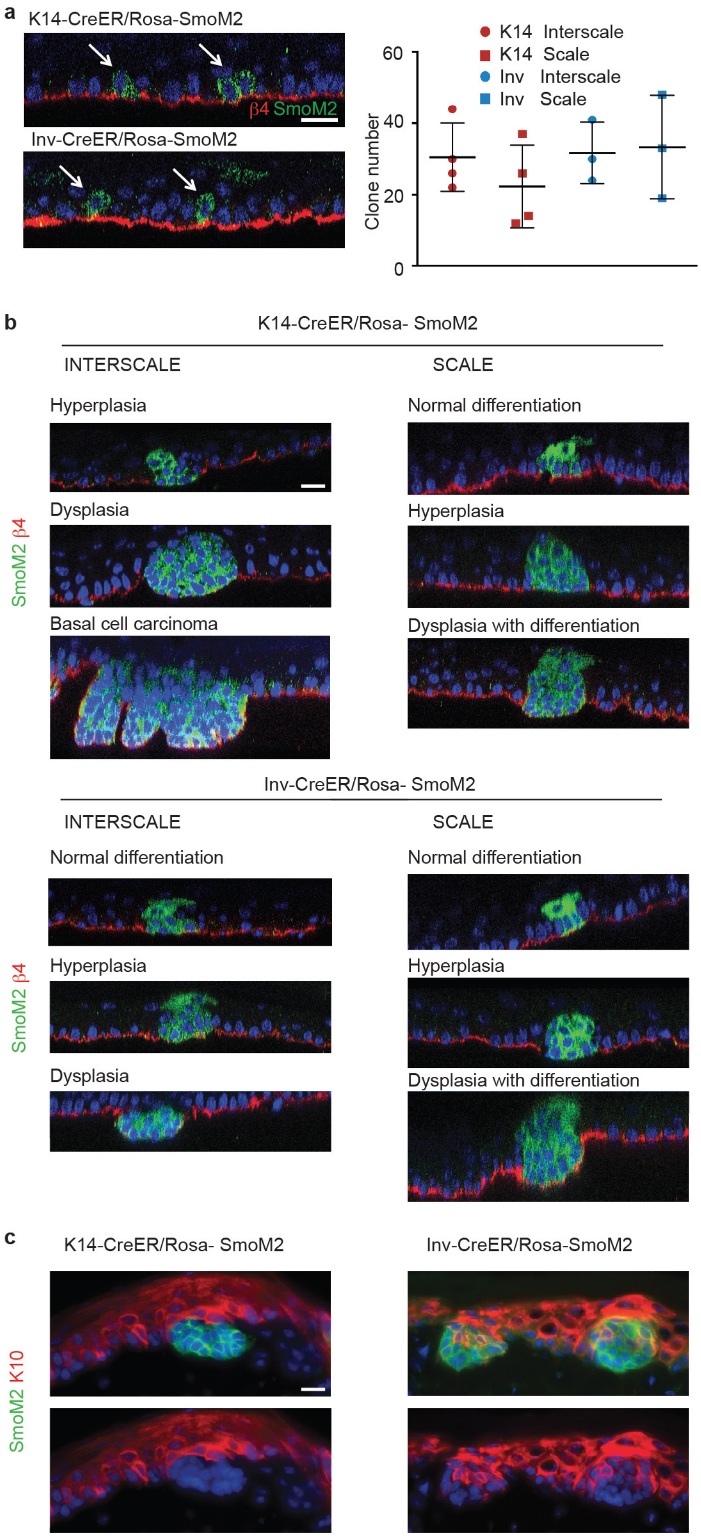
For the analysis of the clone size of K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice (Fig. 5d, g, Extended Data Fig 8e, d) we counted clones from the interscale from two independent experiments at 1, 2 and 4 weeks, three independent experiments at 8 weeks, four independent experiments at 12 weeks. For the analysis of the clone size of Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice (Fig. 5d, g and Extended Data Fig 8e, d) we counted clones from the interscale from two independent experiments at 1 and 2 weeks, three independent experiments at 4 and 8 weeks, four independent experiments at 12 weeks.

For the cell proliferation kinetics experiments in the Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice (Fig. 5f) at 12 weeks after induction 34 clones from 3 independent experiments were counted. For the cell proliferation kinetics experiments in the K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice (Fig. 5f) at 12 weeks after induction 44 clones from two independent experiments were counted.

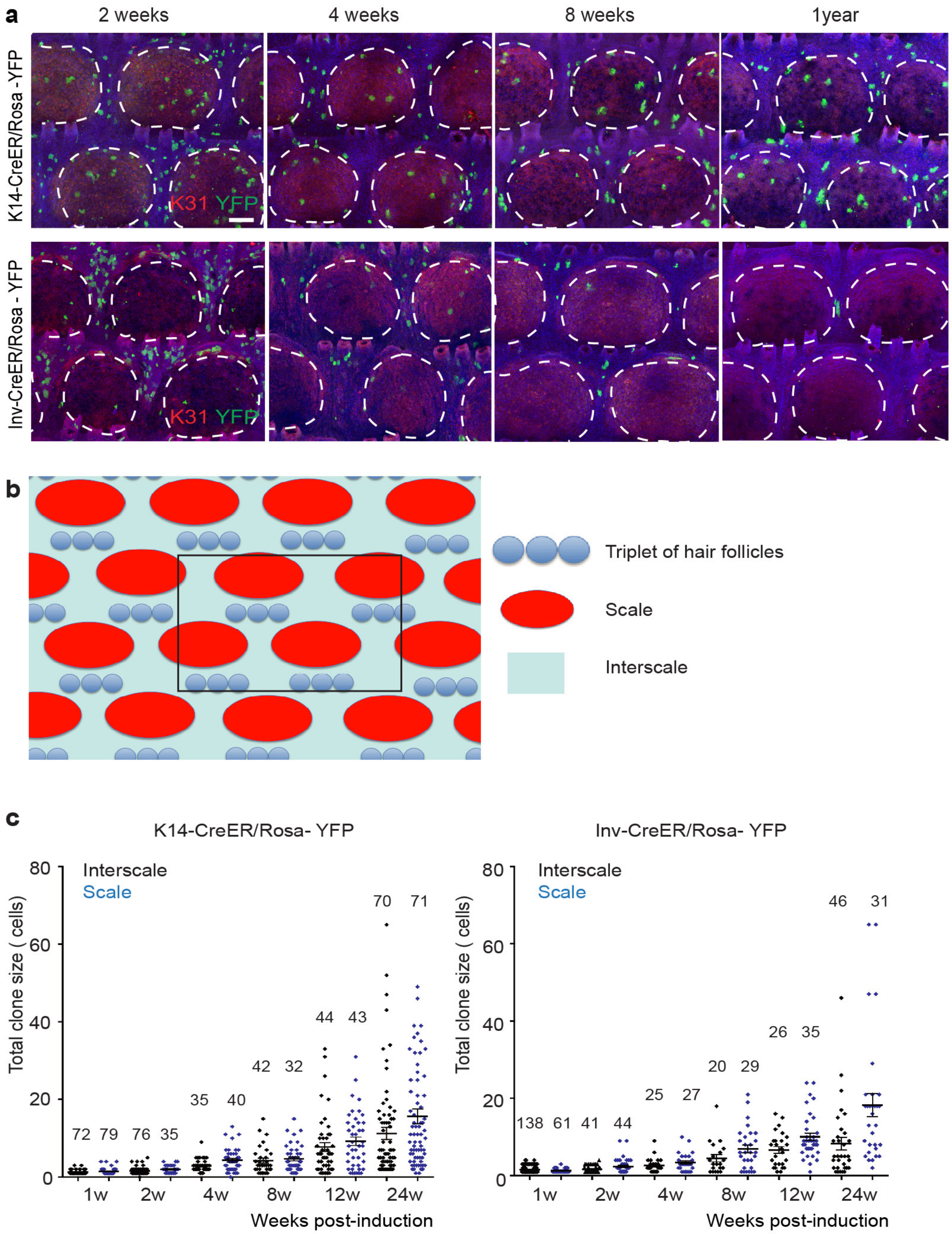
For the clonal persistence experiments in Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$, 132, 78, 68, 58 and 89 clones from 4, 3, 3, 3 and 5 independent experiments were counted at 1, 2, 4, 8, 12 and 24 weeks and in K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice 124, 82, 53, 76 and 100 clones were counted from 4, 3, 2, 3 and 4 independent experiments at 1, 2, 4, 8 and 12 weeks respectively (Extended Data Fig. 8e) (see Source Data).

Source Data includes the persistence and clone size quantifications of K14-CreER/Rosa–YFP, Inv-CreER/Rosa–YFP, K14-CreER/Rosa-SmoM2, Inv-CreER/Rosa-SmoM2, for K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ and Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ animals at different timepoints.

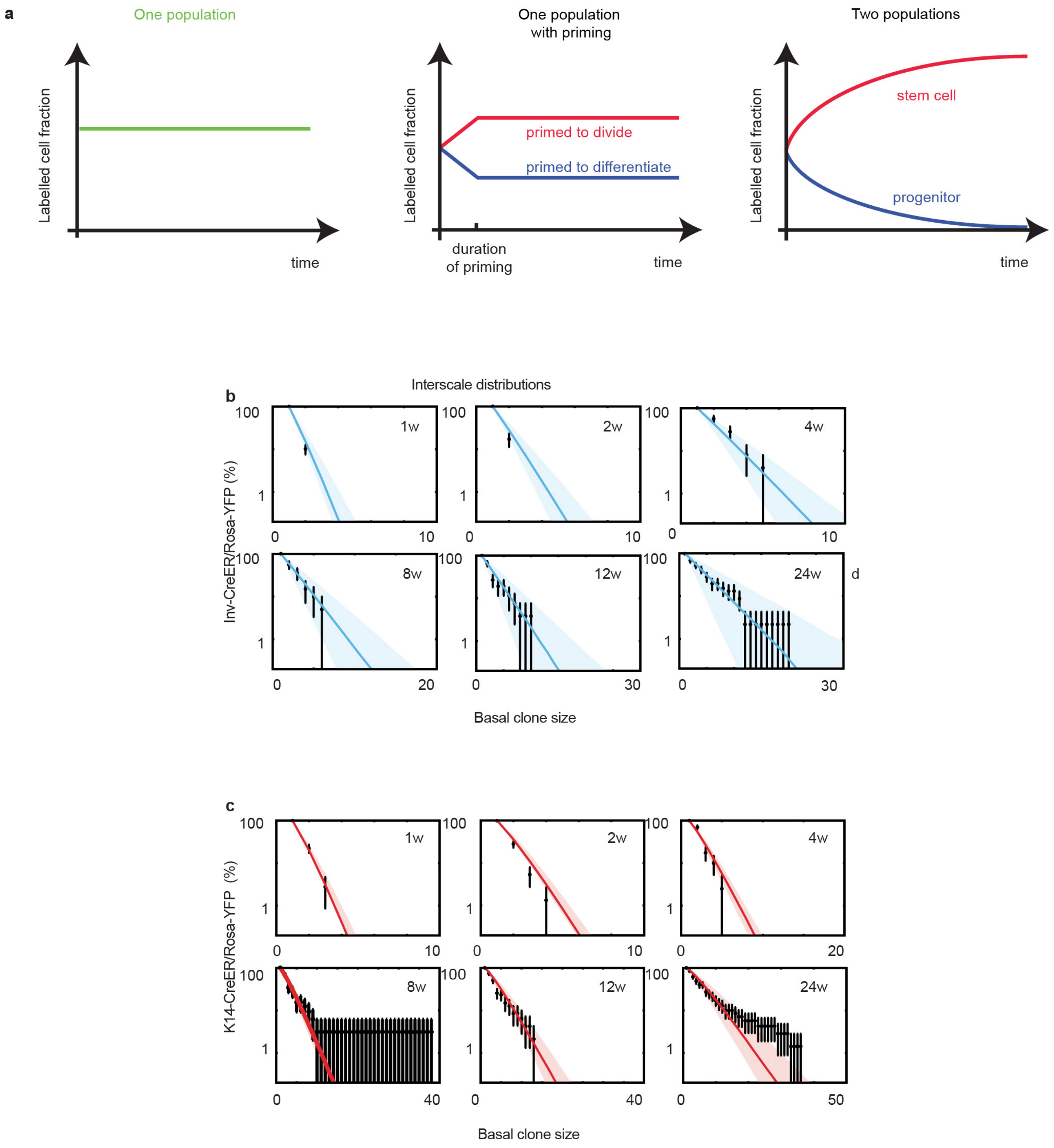
23. Vasioukhin, V., Degenstein, L., Wise, B. & Fuchs, E. The magical touch: genome targeting in epidermal stem cells induced by tamoxifen application to mouse skin. *Proc. Natl Acad. Sci. USA* **96,** 8551–8556 (1999).
24. Uhmann, A. *et al.* The Hedgehog receptor Patched controls lymphoid lineage commitment. *Blood* **110,** 1814–1823 (2007).
25. Mao, J. *et al.* A novel somatic mouse model to survey tumorigenic potential applied to the Hedgehog pathway. *Cancer Res.* **66,** 10171–10178 (2006).
26. Jonkers, J. *et al.* Synergistic tumor suppressor activity of BRCA2 and p53 in a conditional mouse model for breast cancer. *Nat. Genet.* **29,** 418–425 (2001).
27. Braun, K. M. *et al.* Manipulation of stem cell proliferation and lineage commitment: visualisation of label-retaining cells in wholemounts of mouse epidermis. *Development* **130,** 5241–5255 (2003).



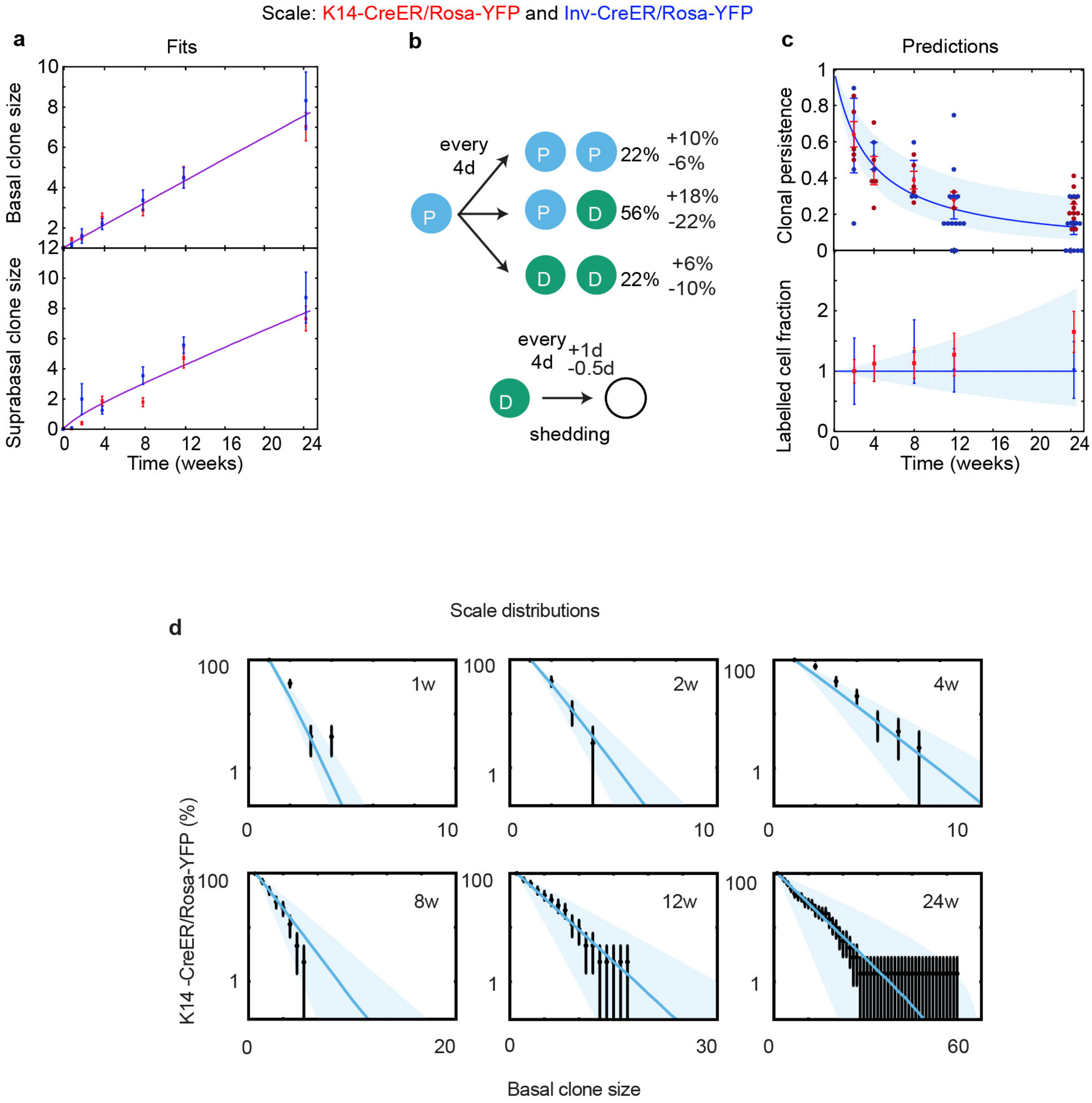
Extended Data Figure 1 | The fate of oncogene-targeted clones is determined by the initial targeted cell (SC or CP) and their location in scale or interscale regions. a, Orthogonal view used to quantify the number of clones, cells stained with β4-integrin and SmoM2. (left). Quantification of the number of clones induced 1 week after tamoxifen administration in scale and interscale regions in K14-CreER/Rosa-SmoM2 ($n = 4$ animals, 0.1 mg tamoxifen) and Inv-CreER/Rosa-SmoM2 ($n = 3$ animals, 2.5 mg tamoxifen) (right). **b**, Immunostaining for β4-integrin and SmoM2 in K14-CreER/Rosa-SmoM2 and Inv-CreER/Rosa-SmoM2 clones located in the scale and interscale regions, 8 weeks after oncogene activation. **c**, Immunostaining for the differentiation marker keratin-10, K10, and SmoM2 in K14-CreER/Rosa-SmoM2 and Inv-CreER/Rosa-SmoM2 clones 8 weeks after oncogene activation, showing absence of differentiated cells in K14-CreER/Rosa-SmoM2 clones and alteration of the differentiation in Inv-CreER/Rosa-SmoM2 clones. Hoechst nuclear staining is represented in blue; scale bars, 10 μm.



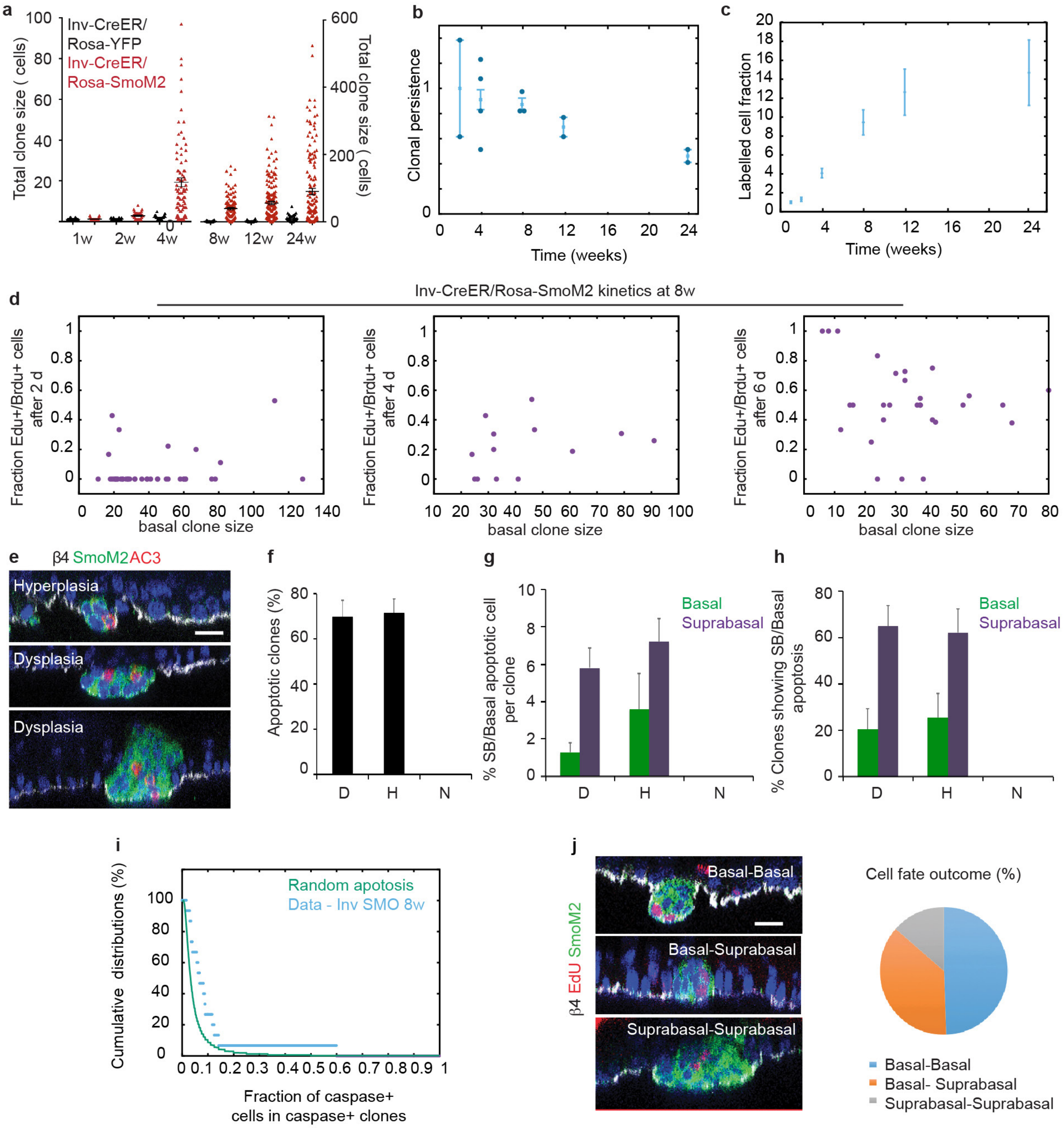
Extended Data Figure 2 | Evolution of K14-CreER/Rosa-YFP and Inv-CreER/Rosa-YFP clones in scale and interscale regions. a, Whole-mount immunostaining for YFP/K31 in K14-CreER/Rosa-YFP mice and Inv-CreER/Rosa-YFP mice upon tamoxifen administration. **b**, Scheme representing the area of tail epidermis (area comprised by 6 groups of triplets of hair follicles, highlighted in black) that is used to quantify the clone number and persistence. **c**, Distribution of K14-CreER/Rosa-YFP and Inv-CreER/Rosa-YFP total clone sizes as measured by total cell content of surviving clones, imaged by confocal microscopy on whole-mount tail epidermis from 1 to 24 weeks after tamoxifen administration. The number of analysed clones is indicated for each time point. Hoechst nuclear staining is represented in blue; scale bars, 100 μm. Histograms and error bars represent the mean and the standard error of the mean (s.e.m.).



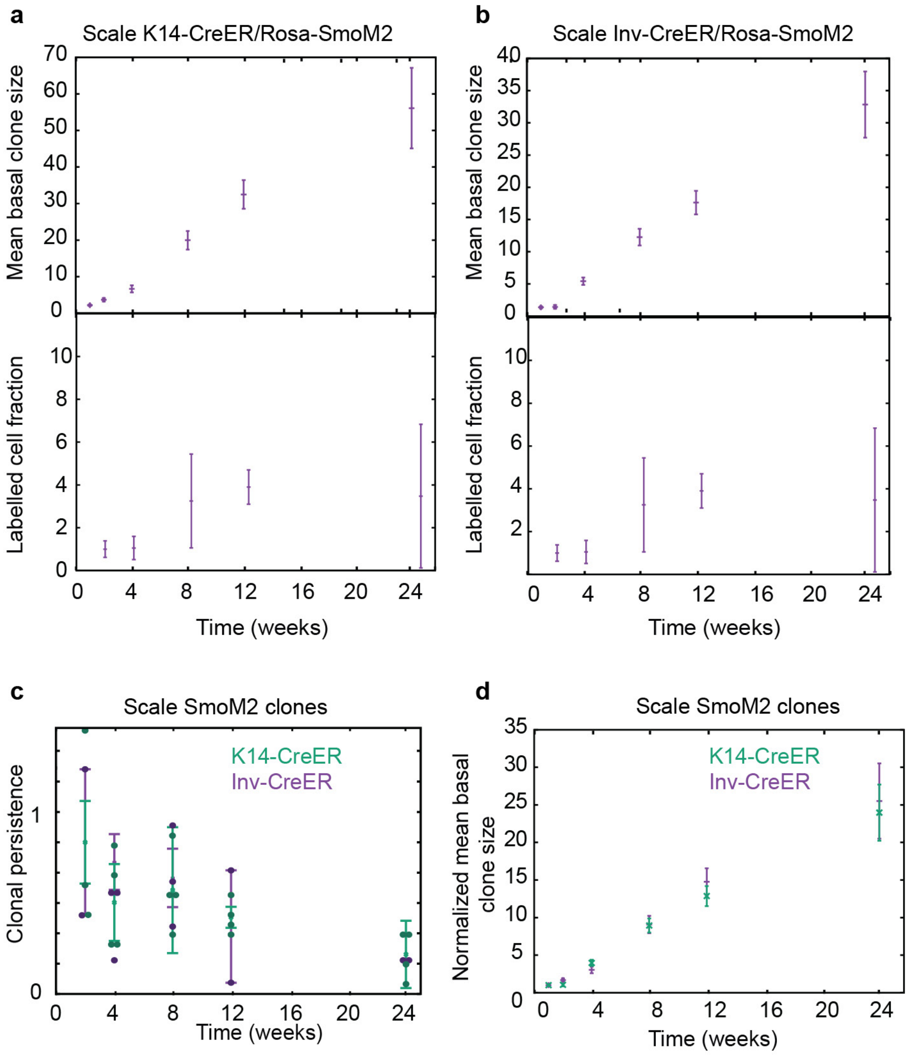
Extended Data Figure 3 | The interscale is maintained by two cell populations during homeostasis. a, Evolution in time of the total labelled cell fraction under three hypotheses. For a perfect single population of equipotent balanced progenitors, the labelled cell fraction remains constant. For a single population of equipotent balanced progenitors displaying short-term priming, the labelled cell fraction increases transiently for the cells primed to divide, and decreases transiently for the cells primed to differentiate, but after the priming period, both fractions remain constant at different values. For two populations organized in a hierarchy, the labelled fraction of the progenitors decreases continuously to zero, while the labelled fraction of the stem cells continuously increases to reach a steady state value, corresponding to its average progeny size. **b**, Cumulative basal clone size distribution of Inv-CreER/Rosa-YFP clones at homeostasis in the interscale upon tamoxifen administration. **c**, Cumulative basal clone size distribution of K14-CreER/Rosa-YFP clones at homeostasis in the interscale upon tamoxifen administration. Clonal distributions are plotted in log-plot, error bars indicate s.d., thick lines are the model prediction and shaded areas indicate 95% confidence intervals in the model prediction.



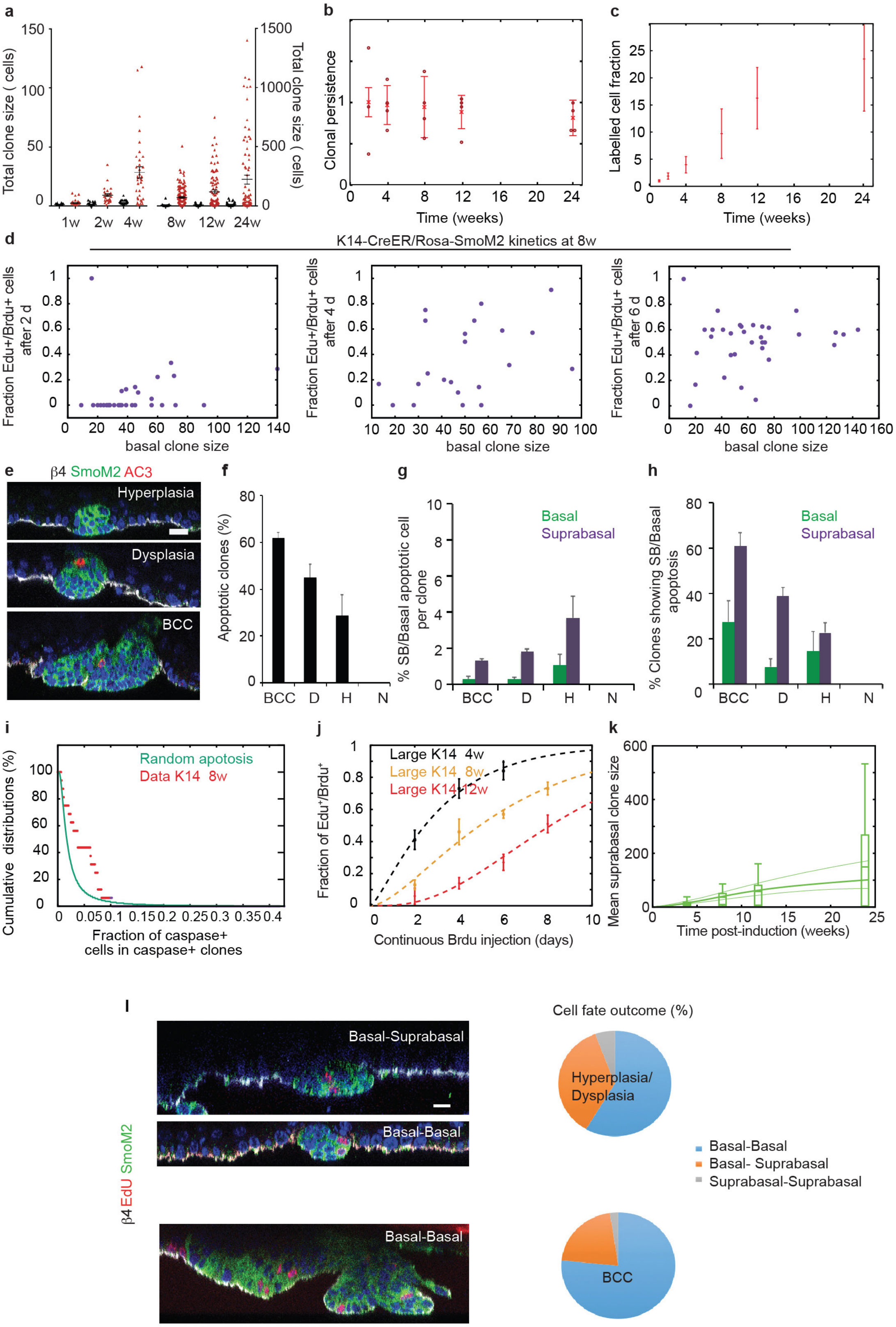
Extended Data Figure 4 | The scale is maintained by a single population during homeostasis. a, Evolution of mean surviving basal (top) and suprabasal (bottom) clone size in the scale for K14-CreER/Rosa-YFP (red) and Inv-CreER/Rosa-YFP (blue). In contrast to the interscale, in the scale K14-CreER and Inv-CreER clones behave identically, indicative of a single progenitor pool. The lines are the fit from the model from which we extract the fate choices of progenitors displayed in **b**. **b**, Fate choices of the equipotent progenitor pool in the scale, as extracted from the fits. **c**, Clonal persistence (top) and labelled cell fraction (bottom) in the scale for K14-CreER/Rosa-YFP (red) and Inv-CreER/Rosa-YFP (blue). The blue and red lines are the predictions of the model (see Supplementary Notes for details) using only the parameters extracted in **b**. K14- and Inv-CreER clones behave similarly and display near-perfect long-term balance. For the clonal persistence data, we examined in each mouse a randomly chosen area shown in Extended Data Fig. 2b. Error bars represent the s.e.m. **d**, Cumulative basal clone size distribution of K14-CreER/Rosa-YFP clones at homeostasis in the scale upon Tamoxifen administration. One should note that there were too few Involucrin clones in the scale to plot meaningful distributions. Clonal distributions are plotted in log-plot, error bars indicate s.d., thick lines are the model prediction and shaded area indicate 95% confidence intervals in the model prediction.



Extended Data Figure 5 | Clonal dynamics of interscale Inv-SmoM2 clones is consistent with a single imbalanced population of progenitors slowing down in time. a, Distribution of Inv-CreER/Rosa-YFP (black) and Inv-CreER/Rosa-SmoM2 (red) clone sizes as measured by total cell content, imaged by confocal microscopy on whole-mount tail epidermis from 1 weeks to 24 weeks following tamoxifen administration. The number of clones analysed in Inv-CreER/Rosa-SmoM2 is indicated in Fig. 3b. The number of clones counted in Inv-CreER/Rosa-YFP is as indicated in Fig. 2b. **b**, Evolution of the clonal persistence for interscale Inv-CreER/Rosa-SmoM2 clones. **c**, Labelled cell fraction for interscale Inv-CreER/Rosa-SmoM2 clones. **d**, Fraction of EdU–BrdU double-labelled cells as a function of basal clone size at 8 weeks for Inv-CreER/Rosa-SmoM2 clones, for 2 (left), 4 (centre) and 6 (right) days of continuous BrdU incorporation. **e**, Immunostaining for β4-integrin, SmoM2 and active-caspase-3 in Inv-CreER/Rosa-SmoM2 clones at 8 weeks after induction. **f**, Percentage of dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones presenting at least one active-caspase positive cell within the clone at 8 weeks after induction ($n = 73$ clones analysed from 4 independent experiments). **g**, Quantification of the number (%) of basal and suprabasal apoptotic cells in dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones 8 weeks after SmoM2 activation. **h**, Percentage of dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones presenting apoptosis in basal and suprabasal compartments 8 w after oncogenic activation. **i**, Cumulative distribution of the fraction of basal apoptosis as a function of basal cell number in an Inv-CreER/Rosa-SmoM2 clone at 8 weeks (data in blue). The green line is the expected theoretical distribution of apoptotic fraction if apoptosis occurred randomly (following a Poisson process), in any clone with the same probability. The data are statistically different from the random theory, showing that apoptosis clusters in certain clones at a given time point. **j**, Short-term fate outcome of progenitors in Inv-CreER/Rosa-SmoM2 clones at 8 weeks, as assessed by using EdU as a clonal marker. We count only cell doublets and classify them as either basal–basal, basal–suprabasal, or suprabasal–suprabasal ($n = 47$ clones from 3 independent experiments). Immunostaining for β4-integrin, EdU and SmoM2 showing the different type of cell fate outcomes found in Inv-CreER/Rosa-SmoM2 clones. Hoechst nuclear staining is represented in blue; scale bars, 10 μm. Histograms and error bars represent the mean and the s.e.m.

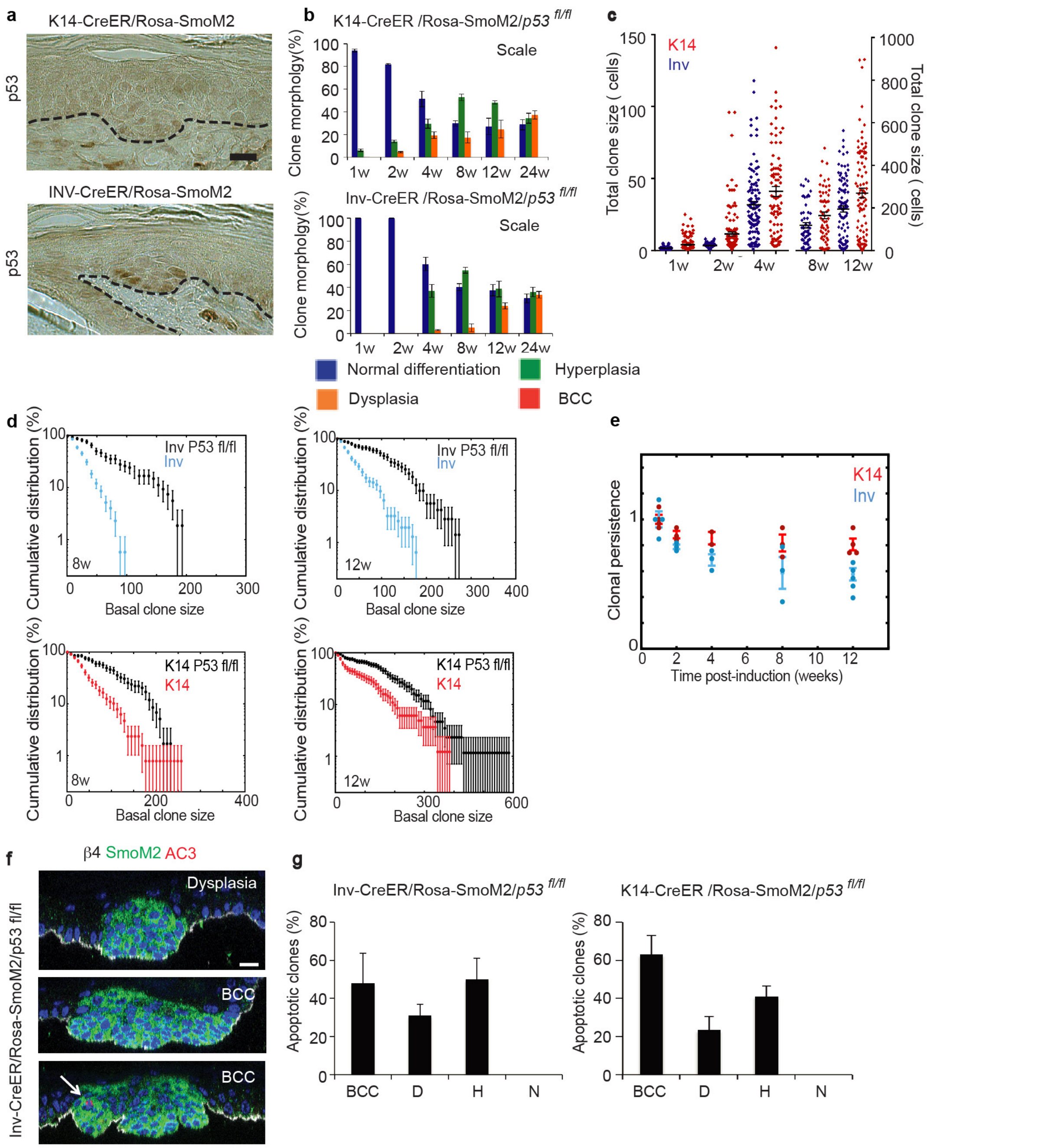


Extended Data Figure 6 | Clonal dynamics of Inv-CreER/Rosa-SmoM2 and K14-CreER/Rosa-SmoM2 clones in the scale are similar. **a**, Evolution of mean surviving basal clone sizes (top) and labelled cell fraction (bottom), for K14-CreER/Rosa-SmoM2, in the scale. **b**, Evolution of mean surviving basal clone sizes (top) and labelled cell fraction (bottom), for Inv-CreER/Rosa-SmoM2, in the scale. Whereas the interscale clones show net expansion, scale clones, both Inv-CreER and K14-CreER, show near balance at the population level. **c**, Evolution of the persistence of K14-CreER/Rosa-SmoM2 (green) and Inv-CreER/Rosa-SmoM2 (purple) clones in the scale. Notably, and in contrast to the interscale, both K14 and Involucrin clones have the same persistence. **d**, Mean basal clone size, normalized by the mean clone size at 1 week for both Inv-CreER and K14-CreER clones. Even though one can see on **a** and **b** that the final clone size is higher in K14, this is fully explained by short-term differences in fate during the first week indicative of short-term priming for K14. Correspondingly, the evolution of the labelling fraction is very similar for K14 and Involucrin in scale. Therefore, K14-CreER/Rosa-SmoM2 and Inv-CreER/Rosa-SmoM2 in scale display the same long-term kinetics upon oncogenic activation, consistent with the one-population model uncovered at homeostasis. Error bars represent the s.e.m.

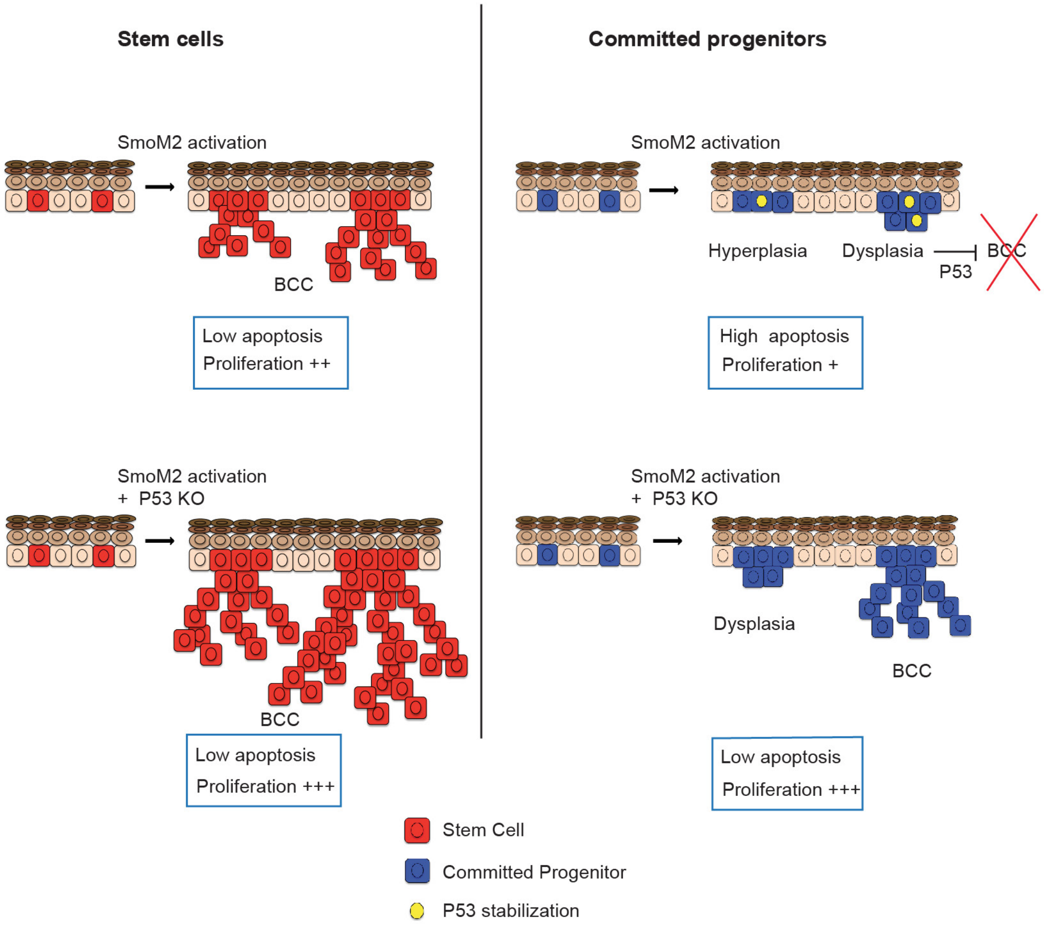


Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Clonal dynamics of interscale K14-CreER/Rosa-SmoM2 clones is consistent with two populations. a, Distribution of K14-CreER/Rosa-YFP (black) and K14-CreER/Rosa-SmoM2 (red) clone sizes as measured by total cell content, imaged by confocal microscopy on whole mount tail epidermis from 1 week to 24 weeks after induction. The number of clones analysed for K14-CreER/Rosa-SmoM2 is indicated in Fig. 4b; the number of clones counted in K14-CreER/Rosa-YFP is as indicated in Fig. 2a. **b**, **c**, Evolution of the clonal persistence (**b**) and labelled cell fraction (**c**) for K14-CreER/Rosa-SmoM2 clones in the interscale. **d**, Fraction of EdU–BrdU double-labelled cells as a function of basal clone size at 8 weeks for K14-CreER/Rosa-SmoM2 clones, for 2 (left), 4 (centre) and 6 (right) days of continuous BrdU incorporation. **e**, Immunostaining for β4-integrin, SmoM2 and active-caspase-3 in K14-CreER/Rosa-SmoM2 clones 8 weeks after SmoM2 activation. **f**, Percentage of BCC, dysplastic, hyperplastic and normally differentiating clones presenting at least one active-caspase-3 positive cell at 8 weeks after induction ($n = 117$ clones analysed from 4 independent experiments). **g**, Quantification of the number (%) of basal and suprabasal apoptotic cells in dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones 8 weeks after SmoM2-activation. **h**, Percentage of dysplastic, hyperplastic and normally differentiating Inv-CreER/Rosa-SmoM2 clones presenting basal and suprabasal apoptosis 8 weeks after oncogenic activation. **i**, Cumulative distribution of the fraction of basal apoptosis as a function of basal cell number in a K14-CreER/Rosa-SmoM2 clone at 8 weeks (data in red). The green line is the expected theoretical distribution of apoptotic fraction if apoptosis occurred randomly (following a Poisson process), in any clone with the same probability. The data are statistically different from the random theory, showing that apoptosis clusters in certain clones at a given time point. **j**, Quantification of EdU–BrdU double-labelled cells as a function of the period of continuous BrdU incorporation for large K14 clones at 4 weeks (black), 8 weeks (orange) and 12 weeks (red) after clonal induction. The dashed lines represent the model fit (Supplementary Theory). **k**, Whisker plot of the suprabasal clone size in the interscale. The boxes delineate the first and third quartiles of the data, and the whiskers delineate the first and last deciles of the data at a given time point. The thick continuous line is the best fit from the model from which we extract the probability of fate choices in tumour SC and progenitors, displayed in Fig. 4g. The thin lines represent the mean clone sizes of SC- (top curve) and CP- (bottom curve) derived clones if they were alone. **l**, Short-term fate outcome of progenitors in K14-CreER/Rosa-SmoM2 clones at 8 weeks, as assessed by using EdU as a clonal marker. We count only cell doublets and classify them as either basal–basal, basal–suprabasal, or suprabasal–suprabasal ($n = 49$ clones from 3 independent experiments). Immunostaining for β4-integrin, EdU and SmoM2 in K14-CreER/Rosa-SmoM2 hyperplastic/dysplastic clones (top) and in BCC (bottom panel). SB, suprabasal. Hoechst nuclear staining is represented in blue; scale bars, 10 μm. Error bars represent the s.e.m.



Extended Data Figure 8 | Effect of p53 deletion in the cellular dynamics of CPs and SCs. a, Immunohistochemistry staining for p53 in Inv-CreER/Rosa-SmoM2 and K14-CreER/Rosa-SmoM2 clones 12 weeks after induction. **b**, Quantification of normal, hyperplastic, dysplastic and BCC clones in scale region of K14CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ and Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ mice. Description of number of counted clones is found in the Methods section. **c**, Distribution of clone sizes as measured by total cell content, imaged by confocal microscopy on whole mount tail epidermis. The number of clones analysed is indicated in Fig. 5d. Clone merger events were observed after 12 weeks following oncogenic activation in K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ preventing the accurate quantification of clonal persistence and clone size at longer times. **d**, Comparison of basal clone size distribution of Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ versus Inv-CreER/Rosa-SmoM2 and K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ versus K14-CreER/Rosa-SmoM2 at 8 weeks and 12 weeks upon tamoxifen administration. **e**, Evolution of the clonal persistence of Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ and K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ clones. **f**, Immunostaining of active-caspase-3 and SmoM2 8 weeks after induction in Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$. **g**, Quantification of the proportion of apoptotic clones in Inv-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ ($n = 90$ clones from 3 independent experiments), and K14-CreER/Rosa-SmoM2/*p53*$^{fl/fl}$ ($n = 82$ animals from 3 independent experiments) 8 weeks after induction. Hoechst nuclear staining is represented in blue; scale bars, 10 μm. Error bars represent the s.e.m.



Extended Data Figure 9 | Model of BCC initiation. Activation of SmoM2 in SCs leads to the generation of BCC owing to an increase in cell proliferation and resistance to apoptosis. However, activation of p53 in SmoM2-expressing CPs restricts the progression of dysplastic clones to BCC by promoting apoptosis and cell-cycle arrest. Deletion of p53 in CPs allows them to progress into BCC.

doi:10.1038/nature19071

Capturing a substrate in an activated RING E3/E2–SUMO complex

Frederick C. Streich Jr[1] & Christopher D. Lima[1,2]

Post-translational protein modification by ubiquitin (Ub) and ubiquitin-like (Ubl) proteins such as small ubiquitin-like modifier (SUMO) regulates processes including protein homeostasis, the DNA damage response, and the cell cycle. Proliferating cell nuclear antigen (PCNA) is modified by Ub or poly-Ub at lysine (Lys)164 after DNA damage to recruit repair factors. Yeast PCNA is modified by SUMO on Lys164 and Lys127 during S-phase to recruit the anti-recombinogenic helicase Srs2. Lys164 modification requires specialized E2/E3 enzyme pairs for SUMO or Ub conjugation. For SUMO, Lys164 modification is strictly dependent on the E3 ligase Siz1, suggesting the E3 alters E2 specificity to promote Lys164 modification. The structural basis for substrate interactions in activated E3/E2–Ub/Ubl complexes remains unclear. Here we report an engineered E2 protein and cross-linking strategies that trap an E3/E2–Ubl/substrate complex for structure determination, illustrating how an E3 can bypass E2 specificity to force-feed a substrate lysine into the E2 active site.

Ub and Ubl proteins are conjugated to substrate proteins by dedicated three-enzyme cascades involving E1 activating enzymes, E2 conjugating enzymes, and E3 ligases (reviewed in refs 1–4). E1 enzymes catalyse Ubl activation and thioester transfer to E2 enzymes, and E3 Ubl isopeptide ligases often complete the cascade by co-localizing substrates and E2–Ubl complexes to promote bond formation between the Ubl carboxy (C) terminus and substrate (typically lysines).

E2 enzymes can exhibit substrate specificity. The E2$_{\text{Ubc9}}$ catalyses SUMO conjugation to lysine residues in SUMO consensus motifs (first exemplified for Ψ-K-X-E, where Ψ is a hydrophobic residue and K is lysine)[5]. Structures of a consensus site lysine bound to E2$_{\text{Ubc9}}$ revealed E2 residues that contribute to lysine recognition and pK_a suppression to promote catalysis[6–8]. Similarly, important residues within other E2 active sites also contribute to particular E2–lysine specificities[9–14].

Really interesting new gene (RING) domains, and their structural homologues, are found in several hundred proteins with E3 activity for Ub, SUMO, and Nedd8 conjugation[15]. RING and some non-RING E3 enzymes bind the E2–Ubl thioester, stimulating conjugation by organizing the E2–Ubl into a closed activated conformation, first illustrated for a non-RING E3 (ref. 7) and subsequently shown for a variety of RING E3 enzymes[16–22]. Protein inhibitor of activated STATs (PIAS) proteins[23], known as Siz proteins in yeast[24], were discovered in humans as inhibitors of STAT signalling and function in immune and cytokine signalling and cellular regulation. The Siz/PIAS–RING (SP–RING) proteins constitute the largest family of SUMO E3 enzymes, yet available structures lack E2 or substrate[25].

After DNA damage, PCNA is modified by Ub on Lys164 by the E2$_{\text{Rad6}}$/E3$_{\text{Rad18}}$ pair and sometimes extended into polyUb chains by E2$_{\text{Ubc13}}$/UEV$_{\text{Mms2}}$ to recruit repair factors[26–28]. Yeast PCNA is modified by SUMO on Lys164 and Lys127 during S-phase to recruit Srs2 (refs 29–31). Lys164 is a non-consensus lysine that requires the SUMO E3$_{\text{Siz1}}$ for SUMO modification by E2$_{\text{Ubc9}}$. SUMO-modified PCNA enhances Ub modification[32], suggesting pathway cross-talk. As such, PCNA represents a model system for understanding specificity determinants in Ub and SUMO pathways[27].

Here, we present reconstitution of an E2$_{\text{Ubc9}}$–SUMO thioester mimetic, an active E3$_{\text{Siz1}}$ fragment, and substrate PCNA. The techniques used to engineer the E2–Ubl thioester mimetic leave the E2 active-site cysteine available for cross-linking, generating a bridge between the E2 and PCNA at the endogenous site of Ubl modification with the same number of atoms as the predicted tetrahedral intermediate. We report the crystal structure of this complex at 2.85 Å resolution. The structure and biochemical data reveal molecular determinants of this E3 ligase complex that bypasses E2 specificity to promote modification at PCNA Lys164.

Reconstituting E2$_{\text{Ubc9}}$–SUMO/E3$_{\text{Siz1}}$/PCNA

E2–Ubl thioester mimetics with stable E2–Ubl linkages are needed for structural studies because the E2–Ubl thioester is labile. E2 active site Cys to Lys generates a stable peptide bond[16], but its side chain is longer than the native linkage and it could interfere with substrate interactions. In contrast, E2 Cys to Ser generates an ester linkage, but it is labile when combined with E3 enzymes[9,13,17]. As an alternative, an E2–Ubl thioester mimetic was engineered by substituting lysine for Ala129 in E2$_{\text{Ubc9}}$ near the active site Cys93 ($\text{E2}^{\text{A129K}}_{\text{Ubc9}}$). At physiological pH, E3$_{\text{Siz1}}$ stimulates SUMO conjugation to $\text{E2}^{\text{A129K}}_{\text{Ubc9}}$, but not $\text{E2}^{\text{C93K}}_{\text{Ubc9}}$, consistent with E1-catalysed $\text{E2}^{\text{A129K}}_{\text{Ubc9}}$–SUMO thioester formation followed by Lys129 nucleophilic attack (Extended Data Fig. 1a). $\text{E2}^{\text{A129K}}_{\text{Ubc9}}$–SUMO and $\text{E2}^{\text{C93K}}_{\text{Ubc9}}$–SUMO are competitive inhibitors, with values of the dissociation constant for mimetic binding, K_i, close to the substrate concentration at the half-maximal velocity, K_m, for E2–SUMO thioester interactions with E3$_{\text{Siz1}}$ (Extended Data Fig. 1b and Extended Data Tables 1 and 3). Importantly, Cys93 remains available for cross-linking in the $\text{E2}^{\text{A129K}}_{\text{Ubc9}}$ –SUMO mimetic.

Bismaleimidoethane (BMOE) was used first to cross-link Cys93 in $\text{E2}^{\text{A129K}}_{\text{Ubc9}}$–SUMO to PCNA by replacing Lys164 with cysteine. E2–SUMO–BMOE–PCNA was combined with $\text{E3}^{(167-465)}_{\text{Siz1}}$, and interactions were observed for E3/E2–SUMO–BMOE–PCNA by gel filtration, albeit in multiple peaks (Extended Data Fig. 1c). Our own efforts and previous studies[33,34] suggested a second SUMO molecule (SUMO$^{\text{B}}$) might stabilize the complex through non-covalent interactions with the E2$_{\text{Ubc9}}$ backside given the high affinity measured between SUMO and E2$_{\text{Ubc9}}$ (an apparent dissociation constant, K_d, of 25 ± 4 nM (Extended Data Fig. 1d and Extended Data Table 2)). A second SUMO was provided by fusing SUMO$^{\text{B}}$ to the $\text{E3}^{(167-465)}_{\text{Siz1}}$ C terminus where its position relative to E2$_{\text{Ubc9}}$ and E3$_{\text{Siz1}}$ appeared ideal[25]. Using $\text{E3}^{(167-465)}_{\text{Siz1}}$–SUMO$^{\text{B}}$,

[1] Structural Biology Program, Sloan Kettering Institute, 1275 York Avenue, New York, New York 10065, USA. [2] Howard Hughes Medical Institute, 1275 York Avenue, New York, New York 10065, USA.

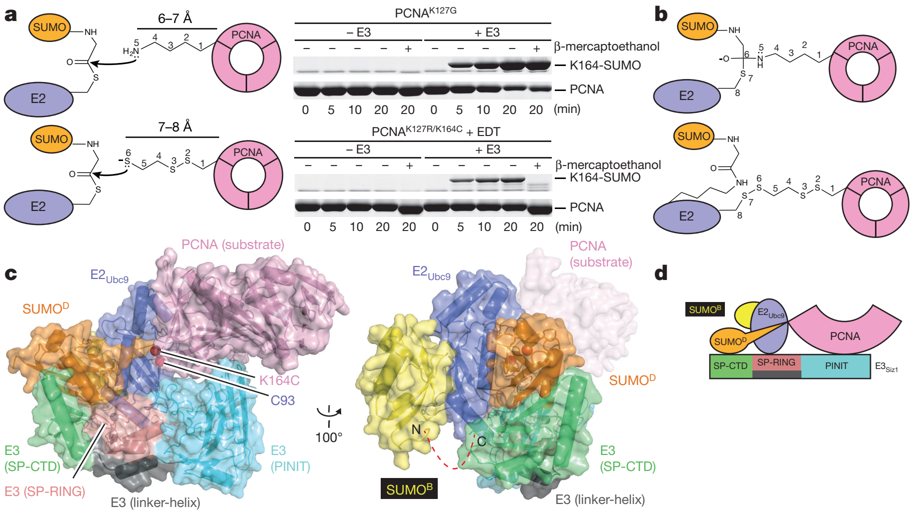


Figure 1 | Reconstituting E2$_{Ubc9}$–SUMO^D/ E3$_{Siz1}$–SUMO^B/PCNA. a, Schematic for nucleophilic attack of the E2–SUMO thioester by lysine or EDT-modified cysteine with *in vitro* SUMO modification of PCNA. **b**, Schematic of the tetrahedral intermediate during transthioesterification and EDT cross-linking of E2$^{A129K}_{Ubc9}$–SUMO and PCNA. **c**, Structure of the complex. **d**, Cartoon model of the complex. For gel source data, see Supplementary Fig. 1.

a stoichiometric complex with E2–SUMO–BMOE–PCNA was observed (Extended Data Fig. 1c). Although BMOE cross-linking trapped the complex, BMOE includes bulky maleimide groups at its ends and it is 4–5 Å longer than the estimated distance spanned by the tetrahedral intermediate (Extended Data Fig. 2a).

1,2-Ethanedithiol (EDT) was identified as a candidate to replace BMOE because it is only one atom longer than lysine when attached to PCNA^{K164C} (Fig. 1a, b). Indeed, PCNA$^{K164C–EDT}$ was a substrate for E3-dependent conjugation by transthioesterification, suggesting EDT can mimic lysine (Fig. 1a and Extended Data Fig. 2b). Furthermore, EDT cross-linking of the E2$_{Ubc9}$ active site cysteine and PCNA^{K164C} yielded a bridge with the same number of atoms between PCNA and the E2 compared with the tetrahedral intermediate (Fig. 1b). E2–SUMO–EDT–PCNA was reconstituted with E3$^{(167–449)}_{Siz1}$–SUMO^B, yielding a monodisperse complex (Extended Data Fig. 2c).

Reconstitutions used trimeric and monomeric PCNA, as both remain dependent on E3$_{Siz1}$ for SUMO modification at Lys164 (ref. 31). Crystals containing trimeric PCNA did not diffract; however, crystals containing monomeric PCNA diffracted to 2.85 Å. The structure was determined and contained two complexes in the asymmetric unit (Fig. 1c, d and Extended Data Table 4). Each complex includes an E2$^{A129K}_{Ubc9}$–SUMO^D thioester mimetic with donor SUMO^D bound to E3$_{Siz1}$ in an activated closed conformation. SUMO^B from E3$^{(167–465)}_{Siz1}$–SUMO^B is bound to the E2$_{Ubc9}$ backside, and EDT bridges Cys93 in E2$^{A129K}_{Ubc9}$ and PCNA^{K164C} above the SUMO^D C terminus which is linked to Lys129 in E2$^{A129K}_{Ubc9}$ (Extended Data Fig. 3a). The SUMO^D C terminus superposes well onto other structures, but its C-terminal carbonyl oxygen points away from E2 Asn85, pushing Cys93 away from the active site (Extended Data Fig. 3a, b). A model of the predicted tetrahedral intermediate requires minimal side-chain movements and no alterations in positions of PCNA relative to E2–SUMO^D (Extended Data Fig. 3c).

E3$_{Siz1}$ SP–RING/SP C-terminal domain activates E2–SUMO^D

The SP–RING domain binds E2$_{Ubc9}$ in a manner similar to E2 interactions with Ub RING domains (Fig. 2a and Extended Data Fig. 3d). Consistent with its function in general activation of E2–SUMO^D, mutations in the E2/E3 interface including Siz1^{I363A}, Siz1^{W387A} and Siz1^{S391D} diminished conjugation to consensus and non-consensus lysine residues[25,33].

The SP C-terminal domain (SP-CTD) was required for SP–RING domain activity, but it was unclear how it worked[25]. Unexpectedly, a SUMO interaction motif (SIM)-like element embedded within the SP-CTD supports binding of SUMO in its activated conformation (Fig. 2b), similar to other activated E2–Ubl complexes (Extended Data Fig. 3e). The SP-CTD is integral to the catalytic module as mutations that disrupt the interface diminished conjugation to consensus and non-consensus lysines (Fig. 2c and Extended Data Fig. 3f)[25]. SIMs usually include three or four hydrophobic amino acids bordered by acidic residues, with the hydrophobic amino acids centred in a β-strand contacting SUMO (reviewed in ref. 2). A hydrophobic substitution (T352V) that makes the SP-CTD more SIM-like increased activity. Mutations with no measurable effect included SUMO^{F37A}, SUMO^{A51I}, Siz1^{Y337A}, Siz1^{Y337E}, and Siz1^{Q431A}.

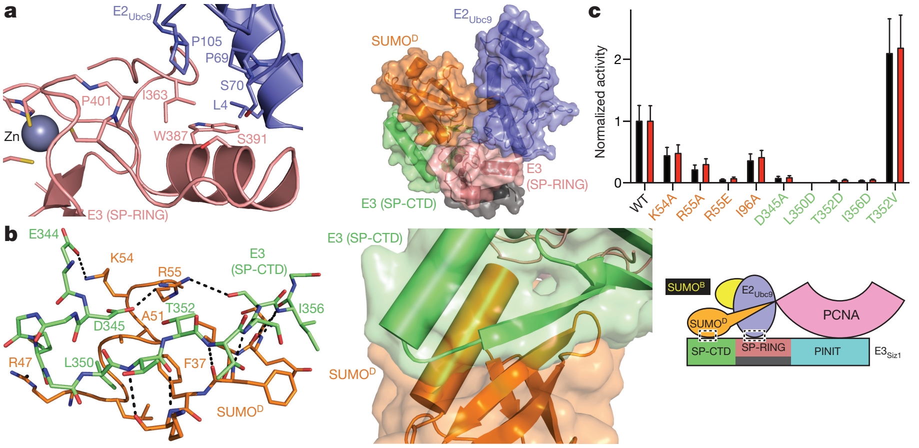


Figure 2 | E3 Activation of E2$_{Ubc9}$–SUMO^D. a, E2/SP–RING interactions (left) and the structure with PINIT removed (right). **b**, SP-CTD/SUMO^D interactions (left) and overview (right). **c**, Quantification of multiple turnover assays of SUMO modification of PCNA with coupled E1, E2, and E3 activities. Quantified rate data show mean ± s.d. ($n = 3$ technical replicates).

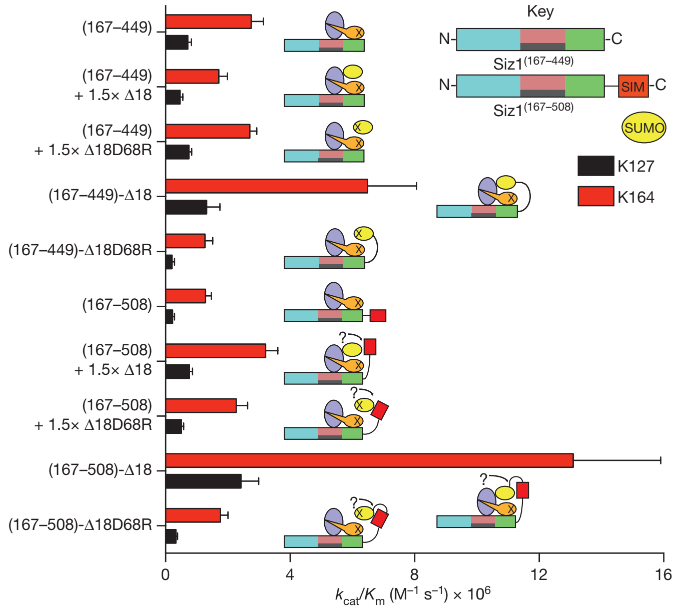


Figure 3 | SUMO^{B} Aids in E2$_{\text{Ubc9}}$–SUMO^{D} Recruitment. Specific activities for E3 and E3–SUMO fusion construct-catalysed multiple turnover reactions with E2–SUMO$^{\text{D68R}}$ thioester titrations with and without 1.5-fold excess SUMO. Quantified rate data show mean ± s.d. ($n = 3$ technical replicates).

SUMO^{B} aids in E2–SUMO^{D} recruitment

SUMO^{B} adopts a similar configuration as observed in other non-covalent E2$_{\text{Ubc9}}$/SUMO^{B} complexes[34–37] (Extended Data Fig. 4a). Several observations suggested that SUMO^{B} could facilitate conjugation. In addition to the high affinity measured between E2$_{\text{Ubc9}}$/SUMO^{B} (Extended Data Fig. 1d and Extended Data Table 2), Siz1 includes a SIM C-terminal to the SP-CTD[38] that could interact with SUMO^{B}. Finally, previous studies implicated E2/SUMO^{B} interactions as important for PIAS and non-RING SUMO E3 activities[33,34]. To evaluate the E3$_{\text{Siz1}}$ SIM and/or SUMO^{B} in conjugation reactions, E3$^{(167-449)}_{\text{Siz1}}$ (no SIM) was compared with E3$^{(167-508)}_{\text{Siz1}}$ (plus SIM, residues 482–486) in the absence or presence of non-conjugatable SUMO or SUMO$^{\text{D68R}}$ at 1.5-fold molar excess, or as E3 C-terminal fusions.

Specific activity for E3$^{(167-449)}_{\text{Siz1}}$ (no SIM) decreased slightly with exogenous SUMO and was unaffected by SUMO$^{\text{D68R}}$; however, specific activity for E3$^{(167-508)}_{\text{Siz1}}$ (plus SIM) increased 2.6-fold with SUMO or 1.8-fold with SUMO$^{\text{D68R}}$ (Fig. 3, Extended Data Fig. 4b and Extended Data Table 3). Effects were most evident with Siz1–SUMO^{B} fusions where kinetic data suggested that SUMO fusions increased activity by decreasing K_m rather than increasing the rate constant k_{cat}, mirroring trends observed in the ubiquitin system[21]. These data suggest that SUMO^{B} can enhance activity.

PCNA binding and substrate specificity

The E3$_{\text{Siz1}}$ PINIT domain forms an interface between the E3 and substrate (Fig. 4a). Consistent with our structure and studies showing that the PINIT domain was required for PCNA Lys164 modification[25], mutation of Siz1 Phe299 or Arg202, the PCNA MEH loop (Met188/Glu189/His190), or a combination, reduced or eliminated detectable modification at Lys164, but not Lys127. Because E2$_{\text{Ubc9}}$ cannot modify PCNA Lys164, we posit that PINIT/PCNA interactions are required to force Lys164 into the E2 active site. Indeed, the modelled conformation for PCNA Lys164 differs from those observed for a SUMO consensus site lysine[6–8,34], Lys63 from Ub[9], or Arg720 from Cullin-1 (ref. 20) (Fig. 4b, c and Extended Data Fig. 3g).

E2$_{\text{Ubc9}}$ side chains near the active site coordinate the lysine nucleophile while lowering its pK_a (ref. 8). We examined whether E3$_{\text{Siz1}}$ makes Lys164 a better nucleophile than Lys127 by differential effects on pK_a suppression; however, single turnover assays revealed no differences (Extended Data Fig. 5a). We next analysed E2 mutations previously implicated in coordinating consensus lysine residues such as Lys127. As anticipated, Y87A disrupted Lys127 modification to below detection; however, Lys164 modification was still evident, albeit diminished (Fig. 4d, e and Extended Data Fig. 5b). In contrast, S127A, S127D, N98A and N124A, selectively reduced activity towards Lys164 compared with Lys127.

Unanticipated interactions were observed between the PINIT FKS loop (residues 268–270) and a loop containing E2$_{\text{Ubc9}}$ Asp100 (Fig. 4b). Specifically, Siz1 Phe268 occupies a hydrophobic pocket on the E2 while backbone nitrogen atoms from Ser270 and Lys269 interact with backbone and side-chain atoms of E2$_{\text{Ubc9}}$ Asp100. Siz1$^{\text{F268A}}$ or deletion of the FKS loop reduced modification at both lysine residues; however, defects were six- to sevenfold greater for Lys164 (Fig. 4d). E2$^{\text{D100A}}_{\text{Ubc9}}$ also decreased modification of Lys164 relative to Lys127. These mutations did not show differential pK_a suppression for

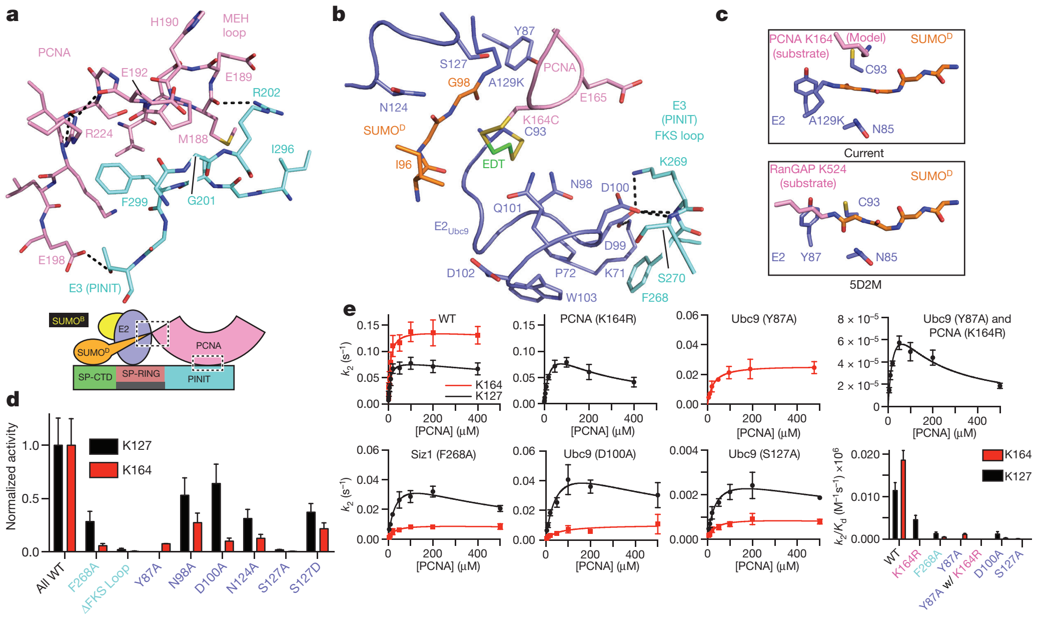


Figure 4 | E3/PCNA interactions and lysine specificity. a, E3$_{\text{Siz1}}$ PINIT/PCNA interactions. **b**, E2 active-site interactions with the FKS loop from the E3$_{\text{Siz1}}$ PINIT domain (EDT in green). **c**, Comparison of the E2$_{\text{Ubc9}}$ active sites with PCNA or RanGAP1. **d**, Quantification of multiple turnover assays for SUMO modification of PCNA with coupled E1, E2, and E3 activities. **e**, Kinetics of single turnover assays with E2$_{\text{Ubc9}}$–SUMO$^{\text{D68R}}$ thioester, E3, and PCNA. For **d** and **e**, quantified data show mean ± s.d. ($n = 3$ technical replicates).

respective lysine nucleophiles (Extended Data Fig. 5a). Owing to its proximity to Lys164, PCNA Glu165 was tested and eliminated as a potential catalytic residue for Lys164 modification (Extended Data Fig. 5c). No measurable defects were observed for $E2_{Ubc9}^{Q101A}$ and $E2_{Ubc9}^{Q101K}$.

Catalytic defects described in preceding sections were explored using single turnover assays to calculate kinetics of modification at each lysine. Apparent K_d values were similar for each lysine using wild-type proteins (~7.5 μM), while the k_2 was 1.6-fold higher for modification of Lys164 over Lys127 (Fig. 4e, Extended Data Fig. 5d and Extended Data Table 5). Selective defects for Lys164 modification by $Siz1^{F268A}$, $E2_{Ubc9}^{D100A}$ and $E2_{Ubc9}^{S127A}$ were attributed to impaired turnover rather than binding. For instance, $Siz1^{F268A}$ reduced k_2 by 1.7-fold for Lys127 versus 15.1-fold for Lys164, while apparent affinities increased 5-fold for Lys127 and 2.8-fold for Lys164. Lys127 modification of wild-type PCNA by $E2_{Ubc9}^{Y87A}$ and $E3_{Siz1}$ was below detection, necessitating use of $PCNA_{K164R}$. While the apparent K_d increased 1.1-fold, a 1,162-fold defect in k_2 was observed for Lys127 modification ($E2_{Ubc9}$ versus $E2_{Ubc9}^{Y87A}$ with $PCNA_{K164R}$). For comparison, the k_2 for Lys164 modification decreased 5.3-fold ($E2_{Ubc9}$ versus $E2_{Ubc9}^{Y87A}$ with wild-type PCNA) while the apparent K_d increased 3.2-fold. Binding defects exist, but mutations altering lysine preference are best explained by differential rate defects, perhaps because these mutants fail to stabilize lysine in productive conformations for catalysis.

Kinetics for Lys127 modification was better modelled by accounting for substrate inhibition while Lys164 modification was better modelled without substrate inhibition (Fig. 4e and Extended Data Table 5). The apparent K_i for wild-type PCNA was 1,430 μM, while the K_i for $PCNA^{K164R}$ was 190 μM. Mutations detrimental for Lys164 modification ($E3_{Siz1}^{F268A}$ and $E2_{Ubc9}^{D100A}$) also resulted in substrate inhibition (K_i of ~400–500 μM). These results suggest that Lys164 is the preferred substrate. When Lys164 modification is impaired, these non-productive complexes behave as inhibitors for Lys127 modification. The high K_i value suggests inhibition is not a factor *in vivo*; however, these results illustrate a complex kinetic relationship between the two modifications.

Surface complementarity between E2–PCNA

PCNA exists as a trimer but our structure includes a monomer of PCNA. Lattice contacts in the crystal reveal PCNA/PCNA contacts reminiscent of the PCNA ring (Extended Data Fig. 6a), and docking trimeric PCNA onto PCNA in our structure revealed no backbone clashes and few side-chain clashes, with $E2_{Ubc9}$ accommodated in the concave surface at the PCNA/PCNA interface (Fig. 5a).

Mutations were generated for $E2_{Ubc9}$ on the penultimate α-helix at positions predicted to point into the PCNA/PCNA interface (Fig. 5b). Each E2 mutant supported conjugation to a SUMO consensus site as evidenced by diSUMO formation (Fig. 5c and Extended Data Fig. 6b, c). In contrast, $E2_{Ubc9}^{E131R}$ resulted in a 44-fold reduction for Lys164 modification and a 2.5-fold increase for Lys127 modification. $E2_{Ubc9}^{R135E}$ and $E2_{Ubc9}^{R139E}$ diminished modification at both sites, although defects were worse for Lys164. $PCNA^{E104R/I100R/L104R}$ or $PCNA^{E113R}$ also diminished conjugation to Lys164 relative to Lys127. In contrast, $E2_{Ubc9}^{K146E/K147E}$, $PCNA^{R61E}$, $PCNA^{T89R}$ and $PCNA^{I91R}$ showed no measurable defect (Extended Data Fig. 6d). These data suggest that steric accommodation of $E2_{Ubc9}$ within the PCNA/PCNA interface contributes to SUMO modification of PCNA at Lys164.

Conclusions

Strategies employed here to link an engineered E2–thioester mimetic and substrate at the endogenous site of modification with the same number of atoms as the predicted tetrahedral intermediate enabled structure determination of an E2–Ubl–substrate/E3 ligase complex, providing a snapshot for otherwise transient states during conjugation. These methods could be used to trap other E2–Ubl/E3/substrate complexes.

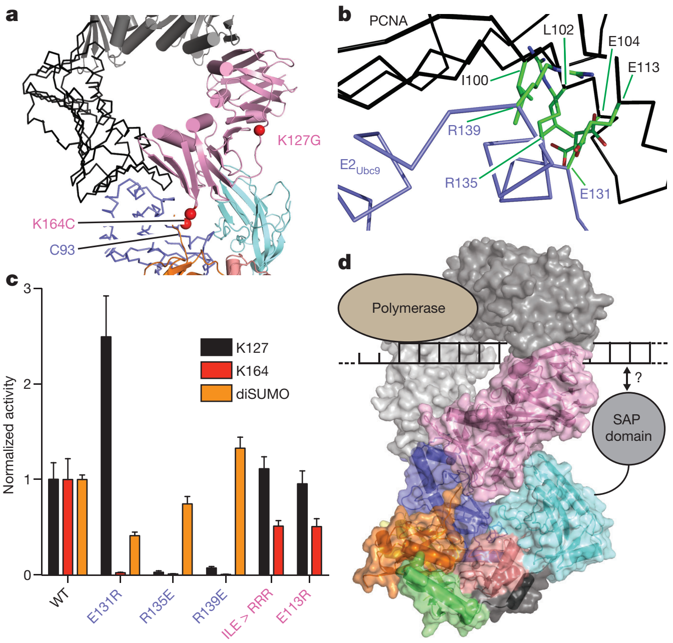


Figure 5 | Surface complementarity between $E2_{Ubc9}$–$SUMO^D$/E3 and PCNA. a, Trimeric PCNA (black ribbon and grey cartoon) modelled onto the current structure (colour). **b**, Location of E2 and PCNA residues with detrimental effects on activity. **c**, Multiple turnover assays of SUMO modification of PCNA with coupled E1, E2, and E3 activities. Quantified data show mean ± s.d. ($n = 3$ technical replicates). **d**, Model of E2–SUMO/E3/trimeric PCNA indicating the site for polymerase opposite the presumed position of the N-terminal $E3_{Siz1}$ SAP domain.

Our structure shows how the SP–RING domain interacts with $E2_{Ubc9}$ and how the SP-CTD coordinates the closed E2–SUMO conformation via an embedded SIM-like motif, extending strategies observed for SIMs in two non-RING SUMO ligases, RanBP2 and ZNF451 (refs 7, 34). Our data expand the utility of a second Ubl ($SUMO^B$) in promoting E2–Ubl/E3 complex formation, including a potential role for the Siz/PIAS SIM motif in recruiting activated $SUMO^B/E2_{Ubc9}$–$SUMO^D$. The affinity measured for $SUMO^B/E2_{Ubc9}$ is ~1,000-fold better than other Ub^B/E2 interactions[21], suggesting $E2_{Ubc9}$ may constitutively bind SUMO *in vivo*. Other possible roles for $SUMO^B/E2_{Ubc9}$ interactions include feed-forward mechanisms, SUMO modification of the E3 could enhance activity as proposed for ZNF451 (ref. 34), or SUMO-modified proteins could amplify conjugation cascades by recruiting SUMO enzymes as proposed in the DNA damage response[39].

E3-dependent modification of PCNA Lys164 does not result because the E3 alters E2-mediated pK_a suppression of Lys164 or Lys127. Instead, specificity arises because the E3 binds the substrate to force-feed Lys164 into the E2 active site. This conformation appears particularly relevant for PCNA Lys164 modification as PCNA cannot be accommodated in the complex on the basis of previously observed substrate lysine conformations (Extended Data Fig. 6e). Furthermore, residues surrounding the E2 active site play differential roles in coordinating Lys164 and Lys127, with a distinct subset of mutations that bias modification of consensus or non-consensus lysine residues (Fig. 4e). It is unclear if $E2_{Rad6}/E3_{Rad18}$ utilize a similar strategy to direct mono Ub modification of PCNA Lys164.

The three-dimensional shape of the E2–Ubl/E3 complex must complement the shape of the substrate for selective modification of a particular lysine(s), a theme noted in other E3 ligase complexes[9,20,40–42]. In our structure, surface complementarity extends beyond E3 contacts to PCNA, and includes complementarity between the E2 and substrate. Although not required *in vitro*, SUMO modification of PCNA is enhanced when PCNA is loaded on DNA, while *in vivo* studies suggest loading of PCNA on DNA is a prerequisite for SUMO modification[43].

Our structure appears consistent with modification of PCNA on DNA as the predicted location of the amino (N)-terminal Siz1 SAP domain that binds duplex DNA is opposite from PCNA surfaces that interact with polymerase (Fig. 5d).

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 April; accepted 30 June 2016.

Published online 10 August 2016.

1. Kerscher, O., Felberbaum, R. & Hochstrasser, M. Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu. Rev. Cell Dev. Biol.* **22,** 159–180 (2006).
2. Gareau, J. R. & Lima, C. D. The SUMO pathway: emerging mechanisms that shape specificity, conjugation and recognition. *Nature Rev. Mol. Cell Biol.* **11,** 861–871 (2010).
3. Hochstrasser, M. Origin and function of ubiquitin-like proteins. *Nature* **458,** 422–429 (2009).
4. Streich, F. C., Jr & Lima, C. D. Structural and functional insights to ubiquitin-like protein conjugation. *Annu. Rev. Biophys.* **43,** 357–379 (2014).
5. Sampson, D. A., Wang, M. & Matunis, M. J. The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. *J. Biol. Chem.* **276,** 21664–21669 (2001).
6. Bernier-Villamor, V., Sampson, D. A., Matunis, M. J. & Lima, C. D. Structural basis for E2-mediated SUMO conjugation revealed by a complex between ubiquitin-conjugating enzyme Ubc9 and RanGAP1. *Cell* **108,** 345–356 (2002).
7. Reverter, D. & Lima, C. D. Insights into E3 ligase activity revealed by a SUMO-RanGAP1-Ubc9-Nup358 complex. *Nature* **435,** 687–692 (2005).
8. Yunus, A. A. & Lima, C. D. Lysine activation and functional analysis of E2-mediated conjugation in the SUMO pathway. *Nature Struct. Mol. Biol*. **13,** 491–499 (2006).
9. Eddins, M. J., Carlile, C. M., Gomez, K. M., Pickart, C. M. & Wolberger, C. Mms2-Ubc13 covalently bound to ubiquitin reveals the structural basis of linkage-specific polyubiquitin chain formation. *Nature Struct. Mol. Biol*. **13,** 915–920 (2006).
10. Bosanac, I. *et al.* Modulation of K11-linkage formation by variable loop residues within UbcH5A. *J. Mol. Biol.* **408,** 420–431 (2011).
11. Wickliffe, K. E., Lorenz, S., Wemmer, D. E., Kuriyan, J. & Rape, M. The mechanism of linkage-specific ubiquitin chain elongation by a single-subunit E2. *Cell* **144,** 769–781 (2011).
12. Saha, A., Lewis, S., Kleiger, G., Kuhlman, B. & Deshaies, R. J. Essential role for ubiquitin-ubiquitin-conjugating enzyme interaction in ubiquitin discharge from Cdc34 to substrate. *Mol. Cell* **42,** 75–83 (2011).
13. Page, R. C., Pruneda, J. N., Amick, J., Klevit, R. E. & Misra, S. Structural insights into the conformation and oligomerization of E2~ubiquitin conjugates. *Biochemistry* **51,** 4175–4187 (2012).
14. Rodrigo-Brenni, M. C., Foster, S. A. & Morgan, D. O. Catalysis of lysine 48-specific ubiquitin chain assembly by residues in E2 and ubiquitin. *Mol. Cell* **39,** 548–559 (2010).
15. Deshaies, R. J. & Joazeiro, C. A. RING domain E3 ubiquitin ligases. *Annu. Rev. Biochem.* **78,** 399–434 (2009).
16. Plechanovová, A., Jaffray, E. G., Tatham, M. H., Naismith, J. H. & Hay, R. T. Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. *Nature* **489,** 115–120 (2012).
17. Dou, H., Buetow, L., Sibbet, G. J., Cameron, K. & Huang, D. T. BIRC7-E2 ubiquitin conjugate structure reveals the mechanism of ubiquitin transfer by a RING dimer. *Nature Struct. Mol. Biol*. **19,** 876–883 (2012).
18. Pruneda, J. N. *et al.* Structure of an E3:E2~Ub complex reveals an allosteric mechanism shared among RING/U-box ligases. *Mol. Cell* **47,** 933–942 (2012).
19. Dou, H., Buetow, L., Sibbet, G. J., Cameron, K. & Huang, D. T. Essentiality of a non-RING element in priming donor ubiquitin for catalysis by a monomeric E3. *Nature Struct. Mol. Biol*. **20,** 982–986 (2013).
20. Scott, D. C. *et al.* Structure of a RING E3 trapped in action reveals ligation mechanism for the ubiquitin-like protein NEDD8. *Cell* **157,** 1671–1684 (2014).
21. Buetow, L. *et al.* Activation of a primed RING E3-E2-ubiquitin complex by non-covalent ubiquitin. *Mol. Cell* **58,** 297–310 (2015).
22. Wright, J. D., Mace, P. D. & Day, C. L. Secondary ubiquitin-RING docking enhances Arkadia and Ark2C E3 ligase activity. *Nature Struct. Mol. Biol*. **23,** 45–52 (2016).
23. Rytinki, M. M., Kaikkonen, S., Pehkonen, P., Jääskeläinen, T. & Palvimo, J. J. PIAS proteins: pleiotropic interactors associated with SUMO. *Cell. Mol. Life Sci.* **66,** 3029–3041 (2009).
24. Johnson, E. S. & Gupta, A. A. An E3-like factor that promotes SUMO conjugation to the yeast septins. *Cell* **106,** 735–744 (2001).
25. Yunus, A. A. & Lima, C. D. Structure of the Siz/PIAS SUMO E3 ligase Siz1 and determinants required for SUMO modification of PCNA. *Mol. Cell* **35,** 669–682 (2009).
26. Hoege, C., Pfander, B., Moldovan, G. L., Pyrowolakis, G. & Jentsch, S. RAD6-dependent DNA repair is linked to modification of PCNA by ubiquitin and SUMO. *Nature* **419,** 135–141 (2002).
27. Moldovan, G. L., Pfander, B. & Jentsch, S. PCNA, the maestro of the replication fork. *Cell* **129,** 665–679 (2007).
28. Parker, J. L. & Ulrich, H. D. Mechanistic analysis of PCNA poly-ubiquitylation by the ubiquitin protein ligases Rad18 and Rad5. *EMBO J.* **28,** 3657–3666 (2009).
29. Pfander, B., Moldovan, G. L., Sacher, M., Hoege, C. & Jentsch, S. SUMO-modified PCNA recruits Srs2 to prevent recombination during S phase. *Nature* **436,** 428–433 (2005).
30. Papouli, E. *et al.* Crosstalk between SUMO and ubiquitin on PCNA is mediated by recruitment of the helicase Srs2p. *Mol. Cell* **19,** 123–133 (2005).
31. Armstrong, A. A., Mohideen, F. & Lima, C. D. Recognition of SUMO-modified PCNA requires tandem receptor motifs in Srs2. *Nature* **483,** 59–63 (2012).
32. Parker, J. L. & Ulrich, H. D. A SUMO-interacting motif activates budding yeast ubiquitin ligase Rad18 towards SUMO-modified PCNA. *Nucleic Acids Res.* **40,** 11380–11388 (2012).
33. Mascle, X. H. *et al.* Identification of a non-covalent ternary complex formed by PIAS1, SUMO1, and UBC9 proteins involved in transcriptional regulation. *J. Biol. Chem.* **288,** 36312–36327 (2013).
34. Cappadocia, L., Pichler, A. & Lima, C. D. Structural basis for catalytic activation by the human ZNF451 SUMO E3 ligase. *Nature Struct. Mol. Biol*. **22,** 968–975 (2015).
35. Knipscheer, P., van Dijk, W. J., Olsen, J. V., Mann, M. & Sixma, T. K. Noncovalent interaction between Ubc9 and SUMO promotes SUMO chain formation. *EMBO J.* **26,** 2797–2807 (2007).
36. Capili, A. D. & Lima, C. D. Structure and analysis of a complex between SUMO and Ubc9 illustrates features of a conserved E2-Ubl interaction. *J. Mol. Biol.* **369,** 608–618 (2007).
37. Duda, D. M. *et al.* Structure of a SUMO-binding-motif mimic bound to Smt3p-Ubc9p: conservation of a non-covalent ubiquitin-like protein-E2 complex as a platform for selective interactions within a SUMO pathway. *J. Mol. Biol.* **369,** 619–630 (2007).
38. Stehmeier, P. & Muller, S. Phospho-regulated SUMO interaction modules connect the SUMO system to CK2 signaling. *Mol. Cell* **33,** 400–409 (2009).
39. Jentsch, S. & Psakhye, I. Control of nuclear activities by substrate-selective and protein-group SUMOylation. *Annu. Rev. Genet.* **47,** 167–186 (2013).
40. Chang, L., Zhang, Z., Yang, J., McLaughlin, S. H. & Barford, D. Atomic structure of the APC/C and its mechanism of protein ubiquitination. *Nature* **522,** 450–454 (2015).
41. McGinty, R. K., Henrici, R. C. & Tan, S. Crystal structure of the PRC1 ubiquitylation module bound to the nucleosome. *Nature* **514,** 591–596 (2014).
42. Mattiroli, F., Uckelmann, M., Sahtoe, D. D., van Dijk, W. J. & Sixma, T. K. The nucleosome acidic patch plays a critical role in RNF168-dependent ubiquitination of histone H2A. *Nature Commun*. **5,** 3291 (2014).
43. Parker, J. L. *et al.* SUMO modification of PCNA is controlled by DNA. *EMBO J.* **27,** 2422–2431 (2008).

Supplementary Information is available in the online version of the paper.

Acknowledgements Research at NE-CAT beamlines was funded by P41 GM103403 (National Institutes of Health/National Institute of General Medical Sciences (NIH/NIGMS)), S10 RR029205 (NIH-ORIP (Office of Research Infrastructure Programs) High-End Shared Instrument grant) at the Advanced Photon Source, a US Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory (DE-AC02-06CH11357). Research was supported in part by GM065872 and GM118080 (NIH/NIGMS, C.D.L.) and P30CA008748 (NIH/ National Cancer Institute). The content is the authors' responsibility and does not represent the official views of the NIH. C.D.L. is a Howard Hughes Medical Institute Investigator.

Author Contributions F.C.S. and C.D.L. executed experiments, data analysis, and manuscript preparation.

Author Information Atomic coordinates and structure factors have been deposited in the Protein Data Bank (PDB) under accession number 5JNE. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to C.D.L. (limac@mskcc.org).

Reviewer Information *Nature* thanks B. Eichman, R. Hay and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Cloning, expression, and purification of recombinant proteins. Expression and purification of mature yeast SUMO (Smt3), Smt3$^{\text{K11C}}$, ΔN18Smt3, yeast E1 (Aos1/Uba2ΔC-term$^{1\text{–}554}$), yeast E2 (Ubc9, wild type, and K153R), yeast E3 active fragments (Siz1$^{(167\text{–}465)}$ and Siz1$^{(167\text{–}508)}$), and tag-free yeast PCNA (wild type, K127G, and K164R) have been described previously[6,7,25,31,44,45]. Point mutants of Smt3 (D68R, R55A, and R55E) were introduced into Smt3$^{\text{K11C}}$ by PCR mutagenesis and expressed and purified as above. Non-conjugatable ΔN18Smt3$^{\text{K19R}\Delta\text{GG97/98LeuGlyHis6Stop}}$ and ΔN18Smt3$^{\text{K19R/D68R}\Delta\text{GG97/98LeuGlyHis6Stop}}$ were generated by PCR amplification and insertion into pET28b with NcoI/XhoI without a stop codon, creating a C-terminal His$_6$ extension. These were purified as the other Smt3 proteins without tag cleavage. The Ubc9$^{\text{C93K}}$ point mutant was generated by PCR mutagenesis and expressed and purified as Ubc9. Other Ubc9 point mutants (C5S/A129K, A129K, Y87A, N98A, D100A, D100E, S127A, S127D, N124A, E131R, R135E, and R139E) were introduced into Ubc9$^{\text{K153R}}$ (to minimize auto-conjugation[46]) by PCR mutagenesis and expressed and purified as Ubc9. The yeast Siz1$^{\text{C361D}}$ mutant was generated to minimize oxidative inactivation of the Siz1 SP–RING domain and has comparable activity to wild type. Aspartic acid is found at the analogous position of *Candida albicans* and *Caenorhabditis elegans* Siz1 orthologues. Siz1$^{(167\text{–}449)}$ was cloned into the NdeI/XhoI sites of pET28b and purified as the Siz1$^{(167\text{–}465)}$. The Siz1$^{(167\text{–}465)}$ point mutants (D345A, F268A, I356D, L350D, T352D, and T352V) were generated by PCR mutagenesis and expressed and purified as Siz1$^{(167\text{–}465)}$. Siz1$^{(167\text{–}465)\Delta\text{FKSLoop}}$ was generated by replacing Siz1$^{(167\text{–}465)}$ amino-acid residues 267–274 with Gly-Ser-Gly by PCR mutagenesis and expressed and purified as Siz1$^{(167\text{–}465)}$. Siz1$^{(167\text{–}449)\text{C361D}}$–ΔN18Smt3, Siz1$^{(167\text{–}449)}$–ΔN18Smt3$^{\text{K19R}\Delta\text{GG97/98-L-E-His6}}$, Siz1$^{(167\text{–}449)}$–ΔN18Smt3$^{\text{K19R/D68R}\Delta\text{GG97/98-L-E-His6}}$, Siz1$^{(167\text{–}508)}$–ΔN18Smt3$^{\text{K19R}\Delta\text{GG97/98-L-E-His6}}$, and Siz1$^{(167\text{–}508)}$–ΔN18Smt3$^{\text{K19R/D68R}\Delta\text{GG97/98-L-E-His6}}$ fusion proteins were generated by inserting the indicated Siz1 sequence into the pSmt3 (ref. 47) vector with BamHI/HindIII and subsequently the indicated ΔN18Smt3 sequences into the HindIII/XhoI sites 3′ of Siz1. These were expressed as N-terminal His$_6$–Smt3 fusions that were cleaved of the N-terminal His$_6$–Smt3 and purified as described previously for His$_6$–Smt3–Siz1$^{(167\text{–}465)}$ (ref. 25). Yeast PCNA$^{\text{K77D/C81E/R110D}}$ and PCNA$^{\text{K77D/C81E/R110D/K127G/K164C}}$ were generated to create monomeric versions of PCNA (residues were identified for ability to induce monomerization) with the latter capable of sulfhydryl based cross-linking at the K164 position. PCNA point mutants (I100R/L102R/E104R, E113R, K127R/K164C, E165A, E165K, E165A/E165K) were generated by PCR mutagenesis. All PCNA constructs were expressed and purified as wild-type PCNA.

The Ubc9$^{\text{C93K}}$–Smt3 thioester mimetic was generated in a reaction containing 20 mM BIS-TRIS propane (pH 10.0), 50 mM NaCl, 10 mM MgCl$_2$, 0.1% Tween-20, 2 mM ATP, 1 μM E1, 25 μM Ubc9$^{\text{C93K}}$, and 200 μM Smt3 for 16 h at 30 °C and purified by Superdex75 and MonoQ. The Ubc9$^{\text{A129K/K153R}}$–Smt3 thioester mimetic was generated in a reaction containing 20 mM BIS-TRIS propane (pH 9.5), 50 mM NaCl, 10 mM MgCl$_2$, 0.1% Tween-20, 2 mM ATP, 1.0 μM E1, 200 μM Ubc9$^{\text{A129K/K153R}}$, and 400 μM Smt3 for 1 h at 30 °C and purified by Superdex75. The Ubc9$^{\text{C5S/A129K/K153R}}$-Δ18Smt3 thioester mimetic was generated in a reaction containing 20 mM BIS-TRIS propane (pH 9.5), 50 mM NaCl, 10 mM MgCl$_2$, 0.1% Tween-20, 2 mM ATP, 0.5 μM E1, 100 μM Ubc9$^{\text{C5S/A129K/K153R}}$, and 200 μM Smt3 for 1 h at 30 °C and purified by Superdex75.

Fluorescence polarization. Smt3$^{\text{K11C}}$ was labelled with Alexa Fluor 488 (hereafter Alexa488) maleimide as recommended by the manufacturer. Smt3$^{\text{K11C}}$–Alexa488, Smt3$^{\text{K11C/D68R}}$–Alexa488, and Ubc9 were buffer exchanged into 20 mM HEPES (pH 7.5), 50 mM NaCl, 0.1% Tween-20, and 1 mM β-mercaptoethanol. Fluorescence polarization was performed at 22 °C with a SpectraMax M5 (Molecular Devices) microplate reader in 384-well microplates. The 20 μl incubations contained 50 nM Smt3$^{\text{K11C}}$–Alexa488 or Smt3$^{\text{K11C/D68R}}$–Alexa488 and buffer alone in the first well followed by a serial dilution of E2$_{\text{Ubc9}}$ from 5 nM to 10 μM, performed in triplicate. Data were analysed in Prism fitted to a single-site binding model accounting for receptor depletion, as described previously[31].

Complex reconstitution and crystallization. Bismaleimidoethane (BMOE, Pierce) crosslinking. The purified Ubc9$^{\text{C5S/A129K/K153R}}$–ΔN18Smt3 thioester mimetic (~800 μl at 493 μM) and yPCNA$^{\text{K77D/C81E/R110D/K127G/K164C}}$ (~480 μl at 3,290 μM) were incubated with 1 mM TCEP (Soltec Ventures, reconstituted in 10 mM HEPES (pH 7.5)) for 15 min at 22 °C. Each protein was desalted into cross-linking buffer (20 mM HEPES (pH 7.0), 200 mM NaCl and 5 mM EDTA). Twelve microlitres of 100 mM BMOE in dimethylsulfoxide (DMSO) was added to the E2–SUMO mimetic (~threefold excess of BMOE) and incubated 3 min at 22 °C and desalted to cross-linking buffer. The E2–Smt3–BMOE was mixed with desalted PCNA (~fourfold excess of PCNA) and incubated 15 min to form Ubc9$^{\text{C5S/A129K/K153R}}$–ΔN18Smt3–BMOE–PCNA$^{\text{K77D/C81E/R110D/K127G/K164C}}$ and was quenched with 1 μl β-mercaptoethanol. Complex was purified on Superdex200 in 20 mM TRIS-HCl (pH 8.0), 350 mM NaCl, and 1 mM β-mercaptoethanol, concentrated, adjusted to 200 mM NaCl and purified on MonoQ.

1,2-Ethanedithiol (EDT, Sigma) crosslinking. The purified Ubc9$^{\text{C5S/A129K/K153R}}$–ΔN18Smt3 thioester mimetic (~400 μl at 1,238 μM) and yPCNA$^{\text{K77D/C81E/R110D/K127G/K164C}}$ (~825 μl at 1,575 μM) were incubated with 1 mM TCEP for 15 min at 22 °C. Both were exchanged to cross-linking buffer. Four microlitres of 0.3 M Aldrithiol-2 (AT2, Sigma) in DMSO was added to the desalted E2–ΔN18Smt3, incubated 15 min at 22 °C, and desalted to cross-linking buffer. Five microlitres of 0.3 M EDT in DMSO was added to the desalted E2–ΔN18Smt3–AT2, incubated 15 min at 22 °C, and desalted to cross-linking buffer. Five microlitres of 0.3 M AT2 was added to the desalted E2–ΔN18Smt3–EDT, incubated 15 min at 22 °C, and desalted to cross-linking buffer. The desalted E2–ΔN18Smt3–EDT–AT2 was mixed with the desalted PCNA (~threefold excess of PCNA) and incubated 20 min at 22 °C to form Ubc9$^{\text{C5S/A129K/K153R}}$–ΔN18Smt3–EDT–PCNA$^{\text{K77D/C81E/R110D/K127G/K164C}}$. The complex was purified by Superdex200 and MonoQ chromatography as for BMOE complex, excluding β-mercaptoethanol.

To reconstitute the final complex for crystallization, the purified Ubc9–ΔN18Smt3–EDT–PCNA complex was mixed with the purified Siz1$^{(167\text{–}449)\text{C361D}}$–ΔN18Smt3 fusion in 1:1 ratio. This was dialysed versus 20 mM TRIS-HCl (pH 8.0), 50 mM NaCl, and 5 mM EDTA and resolved by Superdex200 in same buffer. Peak fractions were concentrated to ~7.5 mg/ml, supplemented with TCEP to 1 mM, aliquoted, and flash frozen at −80 °C for later use. The complex was crystallized at 18 °C by hanging-drop vapour diffusion by mixing 0.5 μl of the complex with 0.5 μl of the reservoir solution containing 0.1 M TRIS-HCl (pH 8.5), 5% PEG 10,000, 0.2 M NaCl, 10% glycerol, and 3% dioxane. Crystals were cryoprotected by gradually increasing the glycerol concentration in the drop by repeated additions of well solution supplemented with 30% glycerol. The crystal was removed from the drop and swiped through another drop of the well solution supplemented with 30% glycerol and then flash cooled in liquid nitrogen. Data were collected at 100 K at the 24-IDE beamline at APS with an ADSC Q315 CCD detector at a 0.979 Å wavelength. Data were indexed, integrated, and scaled with HKL-2000 (ref. 48) to a 2.85 Å resolution. Molecular replacement was performed with Phenix[49] using the crystal structures of Ubc9, Smt3, Siz1, and PCNA (PDB 2EKE, 2EKE, 3ID2, and 1PLQ, respectively) as search models. Refinement was performed with Phenix and model building was performed with Coot[50] and CNS[51,52]. The geometry of the structure was analysed with MolProbity[53]. Of the residues, 96.1% were found in the favoured configuration, with 0.06% Ramachandran outliers (one residue). The structure had a clash score of 2.3 (100th percentile) and a MolProbity score of 1.28 (100th percentile). Figures were prepared with PyMol (http://www.pymol.org/).

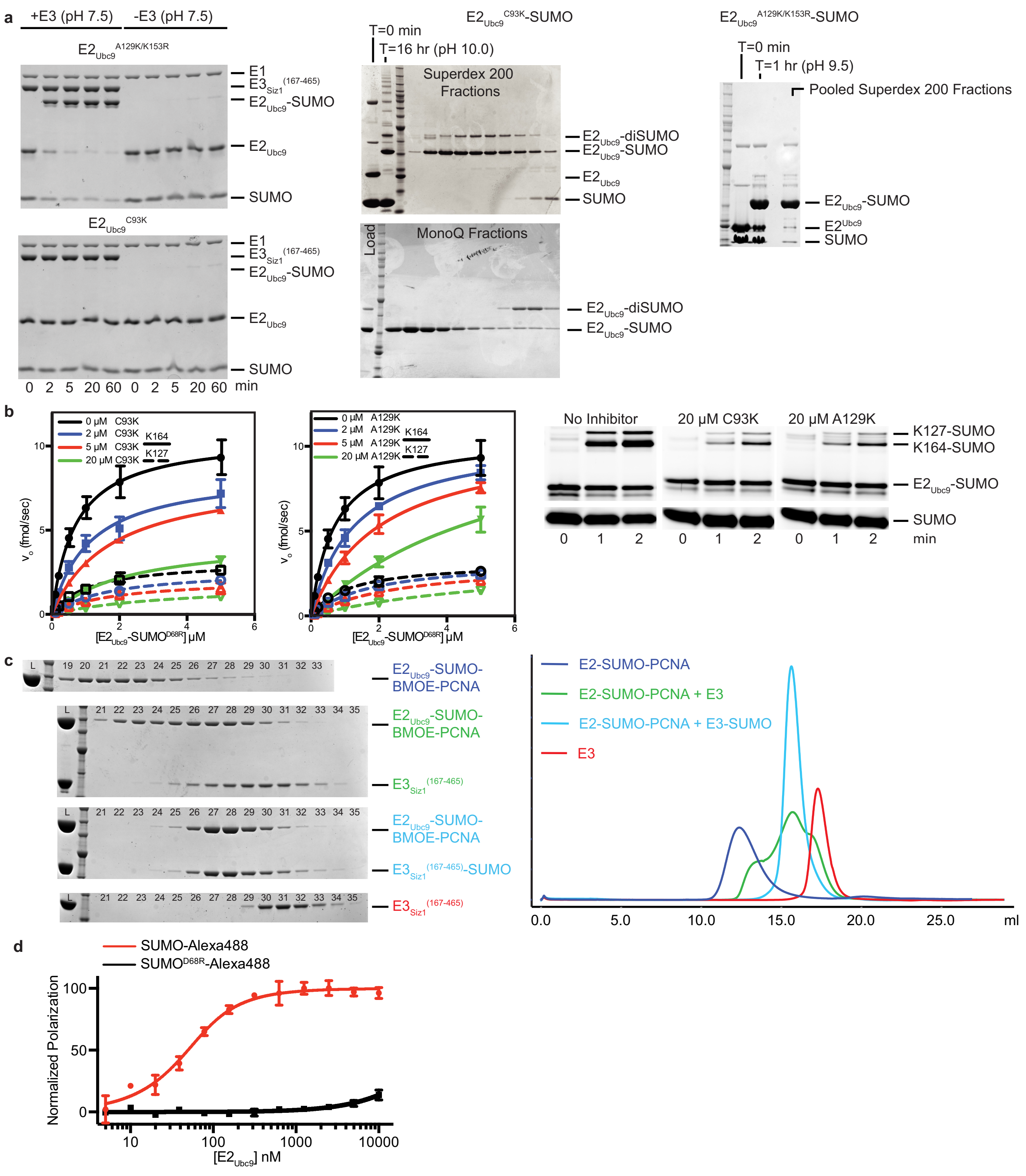
Multiple turnover assay with purified Ubc9$^{\text{K153R}}$–Smt3$^{\text{K11C/D68R}}$–Alexa488 thioester. Smt3$^{\text{K11C/D68R}}$ was labelled with Alexa488 maleimide (Life Technologies) as recommended by the manufacturer. The Ubc9$^{\text{K153R}}$–Smt3$^{\text{K11C/D68R}}$–Alexa488 thioester was formed in a reaction mixture containing 20 mM HEPES (pH 7.5), 50 mM NaCl, 10 mM MgCl$_2$, 0.1% Tween-20, 2 mM ATP, 0.4 mM DTT, 11 μM E1, 200 μM Ubc9$^{\text{K153R}}$, and 100 μM Smt3$^{\text{K11C/D68R}}$–Alexa488 for 5 min at 30 °C. The thioester was diluted and purified by Superdex75 in 50 mM NaCitrate (pH 5.5), 200 mM NaCl, 5% glycerol, concentrated, aliquoted, and flash frozen at −80 °C for later use. A serial dilution of the purified thioester was prepared in 20 mM NaCitrate (pH 5.5), 50 mM NaCl, and 5% glycerol. Ten microlitres of the thioester dilutions were delivered to a 50 μl reaction mixture containing 50 mM HEPES (pH 7.5), 50 mM NaCl, 0.1% Tween-20, 1 nM of the indicated E3, and 32 μM PCNA, and incubated at 30 °C. To determine the K_i for thioester mimetics, parallel reaction mixtures included the thioester mimetic at the concentrations indicated. For the experiments investigating the role of backside-bound SUMO, a 1.5-fold excess of the indicated non-conjugatable ΔN18Smt3 was included in the thioester serial dilution. Aliquots were removed at 1 and 2 min and quenched in equal volumes of 4× LDS NuPAGE loading dye (Life Technologies), resolved by non-reducing 12% SDS–PAGE with MOPS running buffer (Life Technologies), and imaged on Typhoon FLA 9500 with a 473-nm laser and an LPB filter. All gels were imaged with a serial dilution Smt3-Alexa488 reference gel to convert band intensity to picomoles of conjugate with ImageJ (NIH). Experiments were performed in triplicate. Rates were determined by plotting the picomoles of conjugates versus time in Microsoft Excel. Rates were plotted versus E2–thioester concentration in Prism (GraphPad) and fitted to the equation $v = V_{\max}[\text{S}]/(K_{\text{m}} + [\text{S}])$, where $V_{\max} = k_{\text{cat}}[\text{E}]_{\text{t}}$, $[\text{E}]_{\text{t}}$ is the E3 concentration, K_{m} is the substrate concentration at the half-maximal velocity, and [S] is the substrate concentration. K_i was measured by fitting the rate data for all the inhibitor concentrations [I] to the equation $v = V_{\max}[\text{S}]/(K_{\text{m}}(1 + ([\text{I}]/K_{\text{i}})) + [\text{S}])$ in Prism.

Multiple turnover assay with coupled E1, E2, and E3 activities. Reaction mixtures containing 20 mM HEPES (pH 7.5), 50 mM NaCl, 10 mM $MgCl_2$, 0.1% Tween-20, 2 mM ATP, 1 mM DTT, 200 nM E1, 100 nM of indicated $E2_{Ubc9}$, 50 nM of the indicated $Siz1^{(167–465)}$, 80 μM of the indicated Smt3, and 4 μM of the indicated PCNA were incubated at 30 °C. Aliquots were removed at the indicated times and quenched in equal volume of 4× LDS NuPAGE loading dye with 1 M β-mercaptoethanol, resolved by 12% SDS–PAGE with MOPS running buffer. Proteins were stained with SYPRO Ruby (Bio-Rad) and imaged on Typhoon FLA 9500 with a 473-nm laser and an LPG filter. Band intensities were integrated with ImageJ and plotted against time in Microsoft Excel to determine rates relative to wild type. Experiments were performed in triplicate.

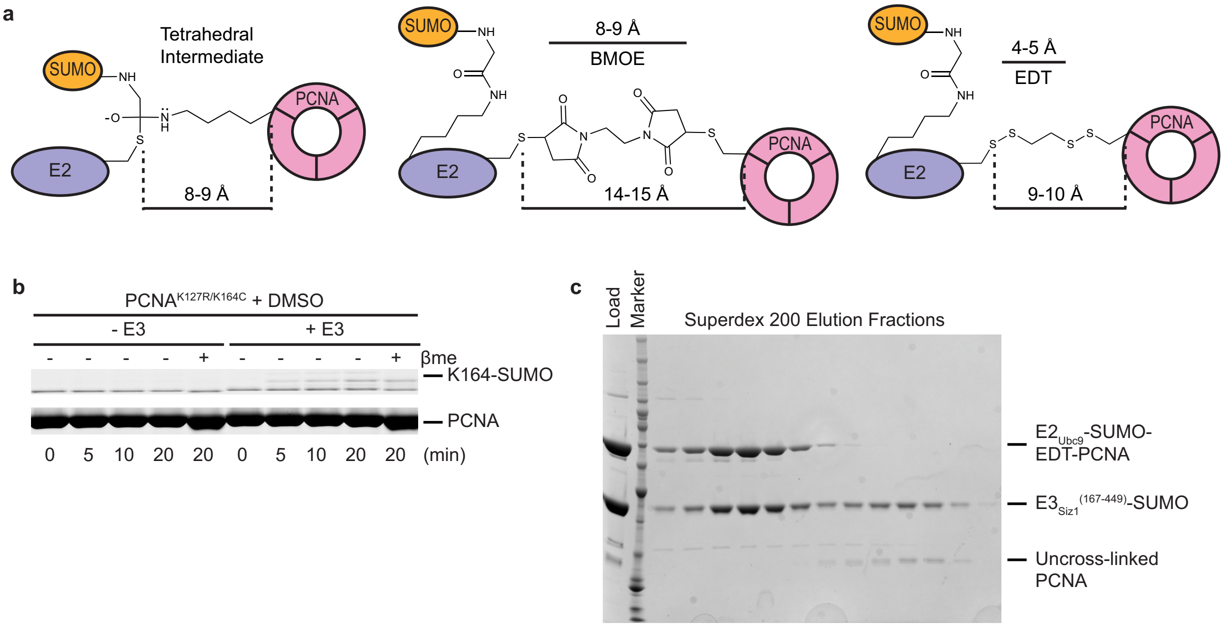
Single turnover assays with purified $Ubc9^{K153R}$–$Smt3^{K11C/D68R}$–Alexa488 and Ubc9 mutant thioesters and Siz1 mutants[25,45]. $Ubc9^{Y87A/K153R}$–$Smt3^{K11C/D68R}$–Alexa488, $Ubc9^{N98A/K153R}$–$Smt3^{K11C/D68R}$–Alexa488, $Ubc9^{D100A/K153R}$–$Smt3^{K11C/D68R}$–Alexa488, $Ubc9^{N124A/K153R}$–$Smt3^{K11C/D68R}$–Alexa488, $Ubc9^{S127A/K153R}$–$Smt3^{K11C/D68R}$–Alexa488 and $Ubc9^{S127D/K153R}$–$Smt3^{K11C/D68R}$–Alexa488 thioesters were formed and purified as above. Ten microlitres of 25 nM of the indicated purified thioester (diluted in 20 mM NaCitrate (pH 5.5), 50 mM NaCl, and 5% glycerol) were delivered to a 50 μl reaction mixture containing 50 mM HEPES (pH 7.5), 50 mM NaCl, 0.1% Tween-20, 50 nM of the indicated E3, and the indicated concentration of PCNA, and incubated at 4 °C. Aliquots were removed at the indicated times, quenched with equal volume of 4× LDS NuPAGE loading dye, resolved by non-reducing 12% SDS–PAGE with MOPS running buffer, and imaged and quantified as above. Experiments were performed in triplicate. Rates were determined by plotting picomoles of conjugates versus time in Microsoft Excel. Owing to the speed of reaction, rates at the higher PCNA concentrations were estimated by the single 5 s time point. Rates were plotted versus PCNA concentration in Prism (GraphPad) and fitted to the equation $v = V_{max}[S]/(K_d + [S])$, where $V_{max} = k_2[E]_t$ for Lys164 data, where $[E]_t$ is E2–Smt3 thioester concentration, K_d is the apparent dissociation constant, and [S] is the substrate concentration. Rates plotted versus PCNA concentration were fitted to the equation $v = V_{max}[S]/(K_d + [S](1 + [S]/K_i))$ for Lys127 data, where $V_{max} = k_2[E]_t$, $[E]_t$ is E2–Smt3 thioester concentration, K_d is the apparent dissociation constant, [S] is the substrate concentration, and K_i is the dissociation constant for substrate binding modelled by considering that two substrates can bind to one enzyme.

Multiple turnover assays with purified E2 thioester at various pH levels. These assays were performed similar to those described for the previous multiple turnover assays, except the reaction and dilution buffers contained BIS-TRIS propane (Sigma) with the pH adjusted from 6.35 to 9.75, measured at 4 °C. The indicated purified E2–thioester was diluted to 700 nM (in 20 mM NaCitrate (pH 5.5), 50 mM NaCl, and 5% glycerol) and 10 μl was delivered to a 70 μl reaction containing 50 mM BIS-TRIS propane (pH as indicated), 50 mM NaCl, 0.1% Tween-20, 5 nM of the indicated $Siz1^{(167–465)}$, and 4 μM PCNA and incubated at 4 °C. Besides the E2–thioester dilution, each protein for each reaction was diluted immediately before initiation to minimize pH effects on protein stability. Aliquots were removed at the indicated times, quenched with equal volume of 4× LDS NuPAGE loading dye, resolved by non-reducing 12% SDS–PAGE with MOPS running buffer, and imaged and quantified as above. Experiments were performed in triplicate.

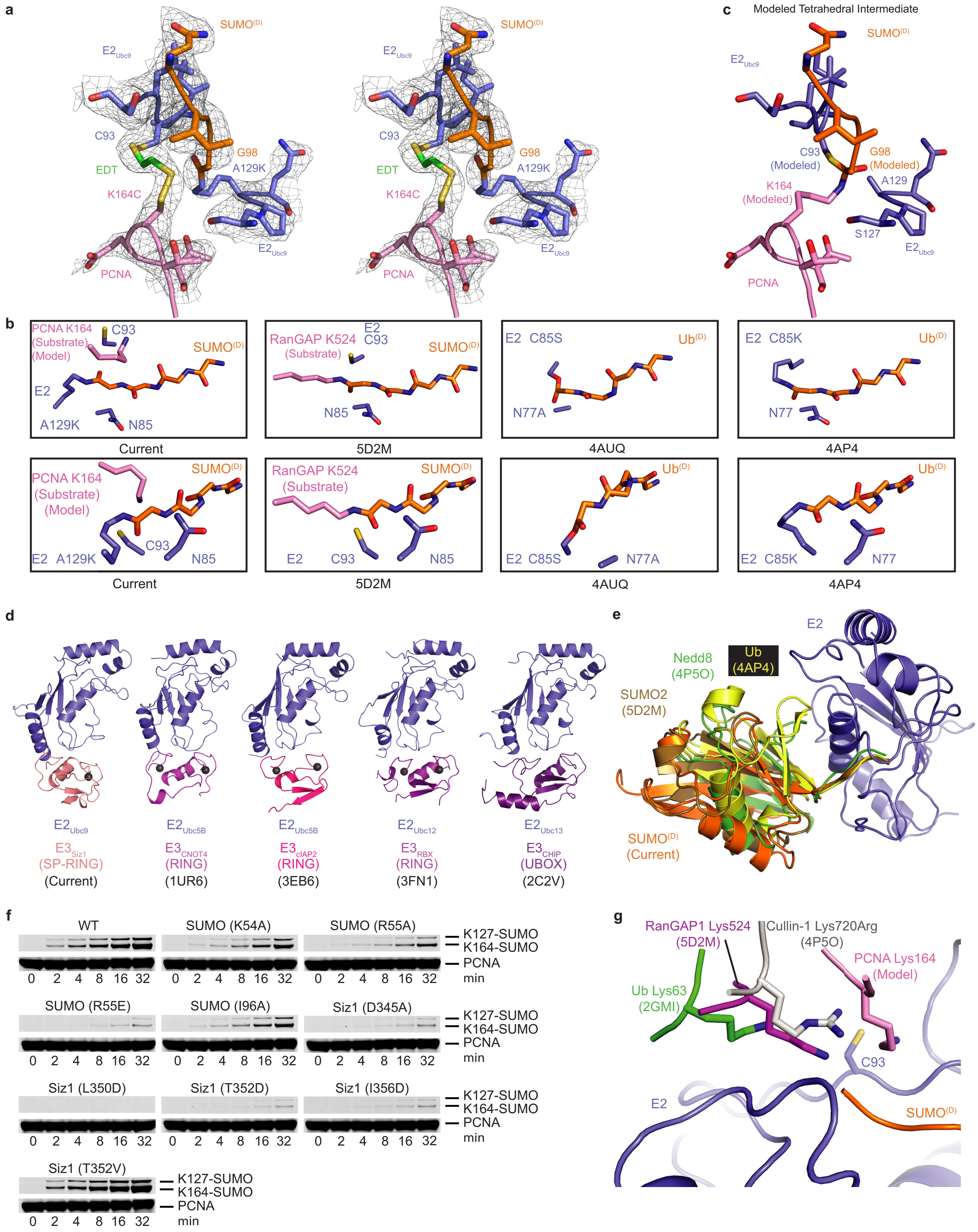
44. Lois, L. M. & Lima, C. D. Structures of the SUMO E1 provide mechanistic insights into SUMO activation and E2 recruitment to E1. *EMBO J.* **24,** 439–451 (2005).
45. Yunus, A. A. & Lima, C. D. Purification of SUMO conjugating enzymes and kinetic analysis of substrate conjugation. *Methods Mol. Biol.* **497,** 167–186 (2009).
46. Knipscheer, P. *et al.* Ubc9 sumoylation regulates SUMO target discrimination. *Mol. Cell* **31,** 371–382 (2008).
47. Mossessova, E. & Lima, C. D. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. *Mol. Cell* **5,** 865–876 (2000).
48. Otwinowski, Z. & Minor, W. in *Methods in Enzymology* Vol. 276 (eds Carter Jr, C. W. & Sweet, R. M.) 307–326 (Academic, 1997).
49. Adams, P. D. *et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
50. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66,** 486–501 (2010).
51. Brünger, A. T. *et al.* Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. D* **54,** 905–921 (1998).
52. Brunger, A. T. Version 1.2 of the Crystallography and NMR system. *Nature Protocols* **2,** 2728–2733 (2007).
53. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).



Extended Data Figure 1 | $E2_{Ubc9}$–SUMO thioester mimic and cross-linking to substrate PCNA for reconstitution with $E3_{Siz1}$. a, SDS–PAGE analysis of *in vitro* $E2^{A129K}_{Ubc9}$ or $E2^{C93K}_{Ubc9}$ charging with SUMO in the presence and absence of $E3^{(167-465)}_{Siz1}$ at pH (7.5) (left) and purification of the $E2^{C93K}_{Ubc9}$–SUMO (middle) and $E2^{A129K}_{Ubc9}$–SUMO (right) thioester mimetics. **b**, Rates for *in vitro* SUMO modification of PCNA in assays using various concentrations of purified $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488-labelled thioester, 1 nM $E3^{(167-465)}_{Siz1}$, and 32 μM PCNA with 0, 2, 5, or 20 μM of the $E2^{C93K}_{Ubc9}$–SUMO or $E2^{A129K}_{Ubc9}$–SUMO thioester mimic (left) with exemplary non-reducing SDS–PAGE for the 0.5 μM $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488 reactions (right). The calculated K_m and K_i from these fits are shown in Extended Data Tables 1 and 3 and the quantified data show mean ± s.d. ($n = 3$ technical replicates). **c**, SDS–PAGE analysis (left) of numbered 0.5 ml fractions from Superose6 analytical gel-filtration analysis (right) of complex reconstitution between $E2_{Ubc9}$–SUMO–BMOE–PCNA and $E3^{(167-465)}_{Siz1}$ (green) or the $E3^{(167-465)}_{Siz1}$–SUMO fusion (blue). Elution profiles for $E2_{Ubc9}$–SUMO–BMOE–PCNA (purple) and $E3^{(167-465)}_{Siz1}$ (red) alone are shown. **d**, The normalized change in polarization observed upon addition of serially diluted $E2_{Ubc9}$ with Alexa488 labelled SUMO or $SUMO^{D68R}$. Data were fitted to a single-site binding model accounting for receptor depletion. Data show mean ± s.d. ($n = 3$ technical replicates). For gel source data, see Supplementary Fig. 1.

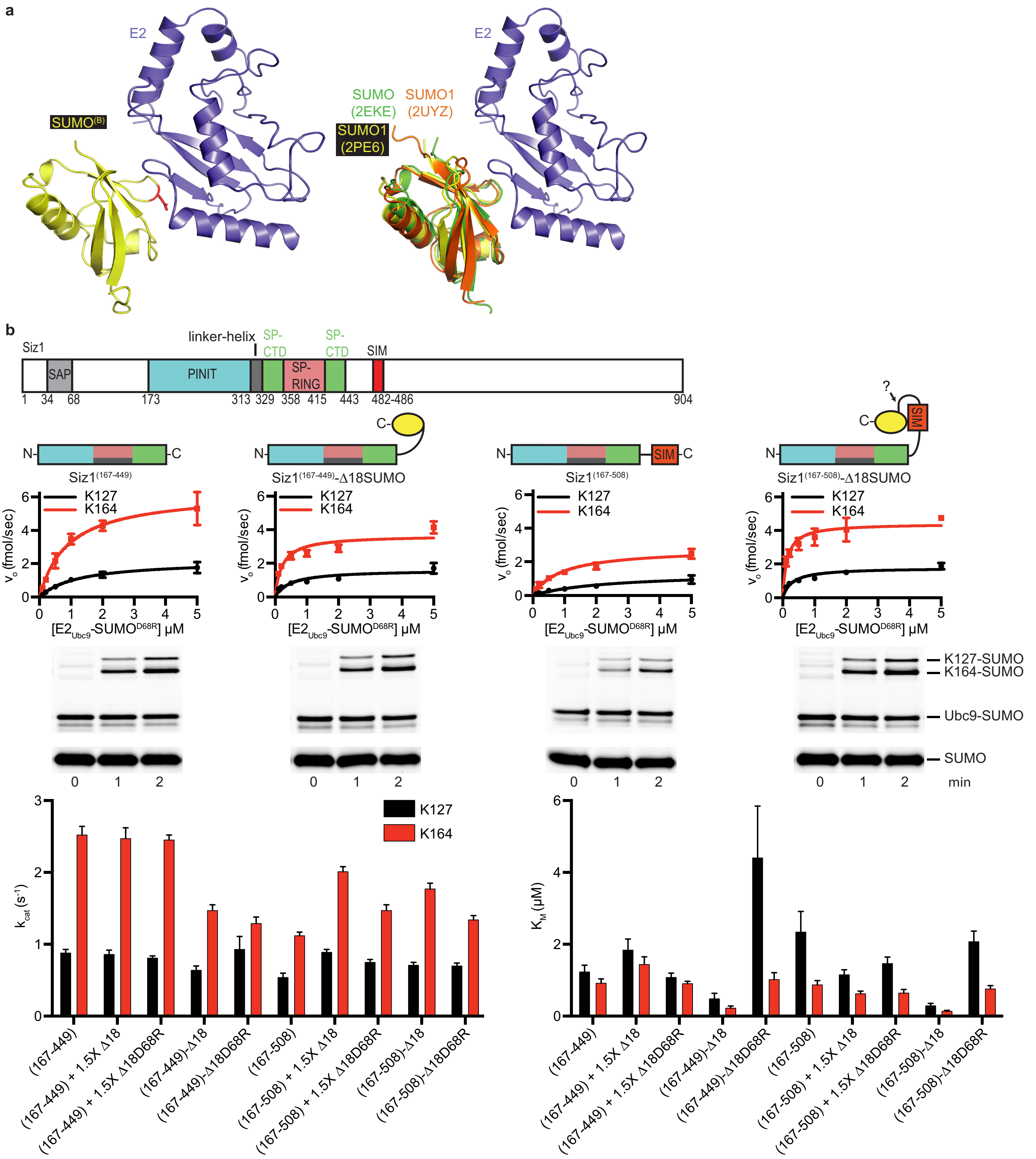


Extended Data Figure 2 | Comparing strategies for crosslinking the E2–SUMO thioester mimic and substrate PCNA. a, Chemical structures of the proposed tetrahedral intermediate formed during PCNA Lys164 attack of $E2_{Ubc9}$–SUMO thioester (left), a BMOE cross-link (middle) or an EDT cross-link (right) between $E2_{Ubc9}$–SUMO C93 and PCNA K164C. Indicated distances were estimated with ChemDraw15 (PerkinElmer). **b**, Control non-reducing SDS–PAGE panel for Fig. 1a showing mock-treated PCNA K127R/K164C (DMSO instead of EDT in DMSO) is unable to accept transthioesterification of SUMO at position 164. **c**, SDS–PAGE analysis of the 5 ml fractions from the final preparative Superdex200 gel-filtration purification of the $E2_{Ubc9}$–SUMO–EDT–PCNA/$E3_{Siz1}^{(167-449)}$–SUMO complex. For gel source data, see Supplementary Fig. 1.

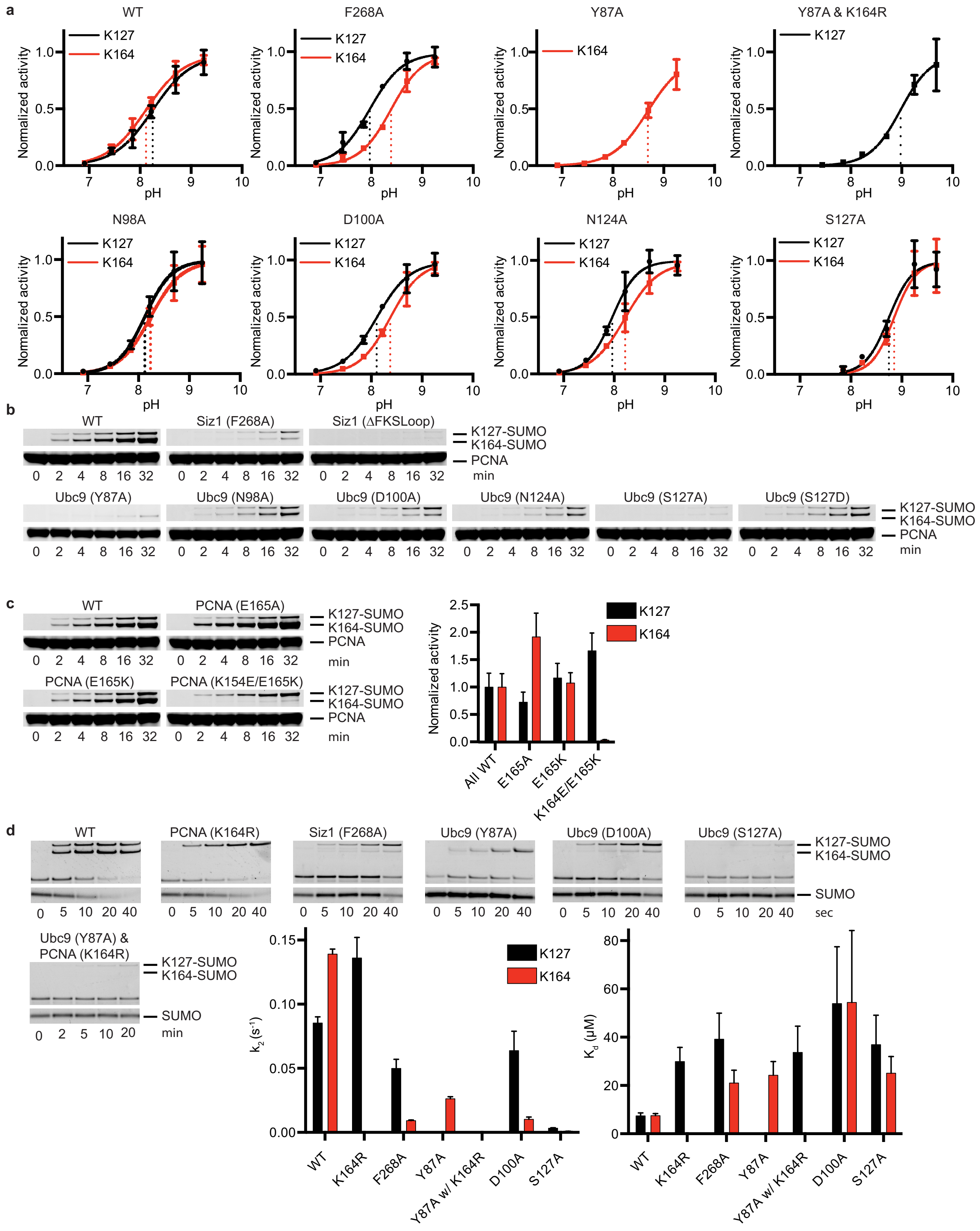


Extended Data Figure 3 | See next page for caption.

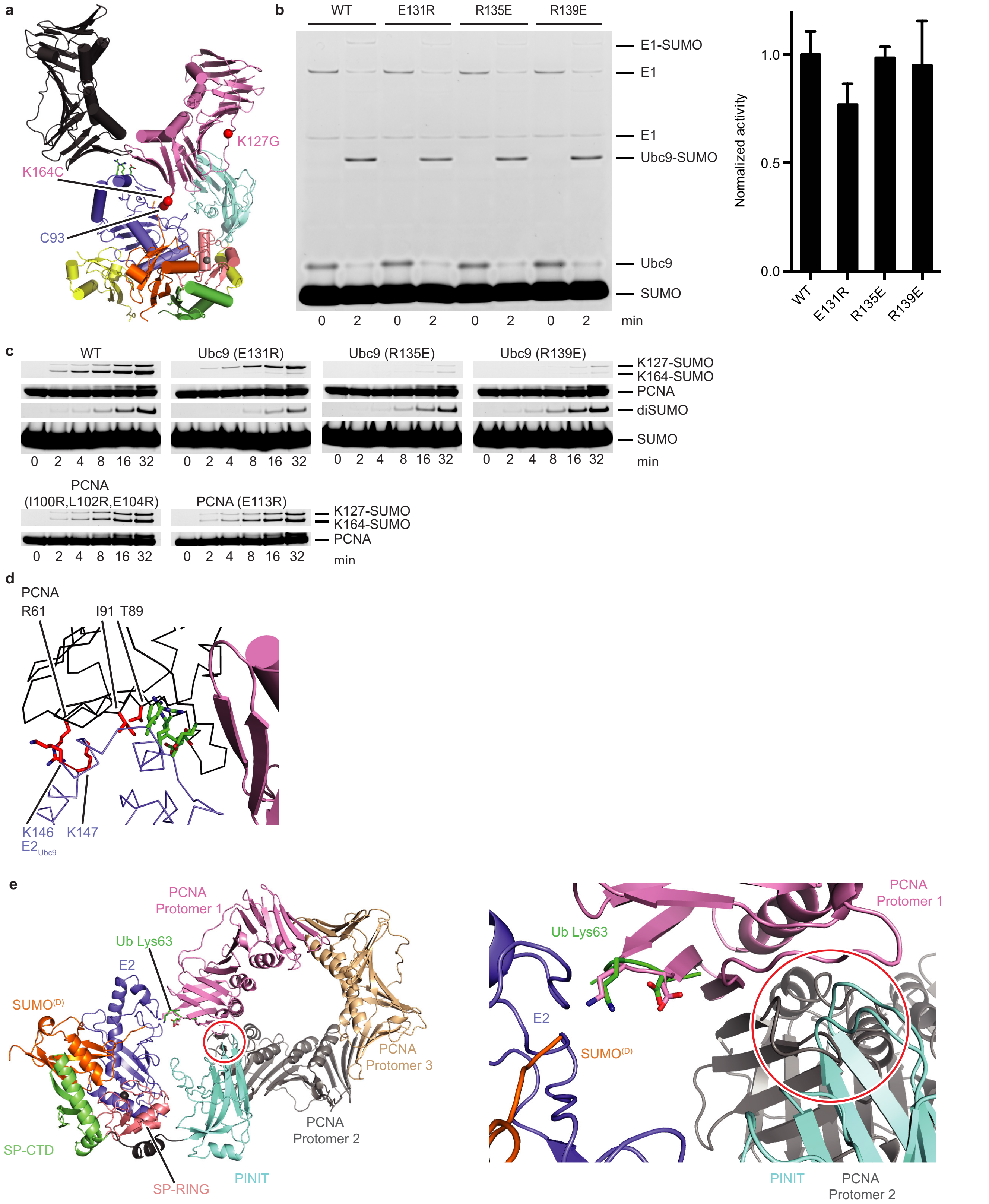
Extended Data Figure 3 | $E2_{Ubc9}$ active site, conformation of $SUMO^{D}$, and comparison with relevant structures. a, Stereo image of simulated annealing electron density map showing the EDT linkage and the SUMO Gly98 linkage to $E2_{Ubc9}$ A129K. The $2F_o - F_c$ electron density map is contoured at 0.8σ (grey mesh). **b**, Alignment of the E2 enzymes from the current structure, SUMO-modified RanGAP1 bound to $E2^{K14R}_{Ubc9}$ and $E3_{Znf451}$ (5D2M), $E2^{S22R/N77A/C85S}_{Ubc5B}$-Ub bound to the RING dimer from $E3_{BIRC7}$ (4AUQ), and E2 $E2^{S22R/C85K}_{Ubc5A}$-Ub bound to the RING dimer from $E3_{RNF4}$ (4AP4) showing two orientations of the E2 active site. **c**, Model of tetrahedral intermediate generated by comparing our structure with other structures of E2–Ubl/E3 complexes, particularly Protein Data Bank (PDB) accession numbers 5DM2 and 4P5O. **d**, Alignment of the current structure and three E2/RING (PDB 1UR6, 3EB6, and 3FN1) complexes and one E2/UBox (PDB 2C2V) complex (aligned by the E2). **e**, Alignments of four E2–Ubl/E3 complexes (aligned by the E2) in the closed activated confirmation for the current structure, $E2^{K14R}_{Ubc9}$–SUMO (PDB 5D2M), $E2^{S22R/C85K}_{Ubc5A}$-Ub (PDB 4AP4), and $E2^{N103S/C111S}_{Ubc12}$-Nedd8 (PDB 4P5O). **f**, SDS–PAGE analysis of multiple turnover assays of SUMO modification of PCNA using *in vitro* reactions with coupled E1 (200 nM), E2 (100 nM), and E3 (50 nM) activities with 4 μM PCNA for the quantified data shown in Fig. 2c. **g**, Alignments of E2 from relevant structures with lysine or arginine residues within or projecting towards the E2 active sites compared with the current structure. Lysine 63 from acceptor ubiquitin projecting towards the active site of the $E2^{C87S}_{Ubc13}$–Ub is shown in green (PDB 2GMI). Lysine 524 from SUMO-modified RanGAP1 laying across the active site of $E2^{K14R}_{Ubc9}$ is shown in magenta. The Lys720Arg from Cullin-1 projecting into the active site of $E2_{Ubc12}$–Nedd8 is shown in grey (PDB 4P5O). For the current structure, EDT was removed from the model, Cys164 was mutated back to lysine, and the side chain was fitted to the electron density and is shown in pink in reference to the current E2 (blue) and donor SUMO (orange). For gel source data, see Supplementary Fig. 1.



Extended Data Figure 4 | SUMO^{B} bound to the E2 backside enhances E2$_{Ubc9}$–SUMO recruitment. a, Alignment of the current E2$_{Ubc9}$/backside SUMO^{B} (left) to previously observed E2$_{Ubc9}$/backside SUMO complexes (right). The position of the D68R mutation is shown in red sticks (left). **b**, Primary E3$_{Siz1}$ structure (top). Cartoons indicating the E3$_{Siz1}$ or E3$_{Siz1}$–SUMO fusion constructs used in the multiple turnover *in vitro* assays (middle) shown in Fig. 3 using a titration of the purified E2$_{Ubc9}$–SUMO^{D68R}–Alexa488 thioester with or without 1.5-fold excess of the indicated additional molecule of non-conjugatable SUMO, 1 nM of the indicated E3 construct, and 32 μM PCNA. Representative non-reducing SDS–PAGE showing the 0.5 μM E2$_{Ubc9}$–SUMO^{D68R}–Alexa488 thioester reactions below the plots of the rates of reaction for each E2$_{Ubc9}$–SUMO^{D68R} concentration (middle). The kinetics of SUMO modification of PCNA were calculated and K_m and k_{cat} determined (bottom); these are shown in Extended Data Table 3. The quantified rate data show mean ± s.d. ($n = 3$ technical replicates). For gel source data, see Supplementary Fig. 1.



Extended Data Figure 5 | $E2_{Ubc9}$ and $E3_{Siz1}$ determinants of lysine specificity. a, Plots of the rates observed at different pH values for multiple turnover *in vitro* assays of SUMO modification of PCNA using 0.1 μM purified $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488 thioester (or $E2_{Ubc9}$ mutant thioesters) with 5 nM $E3_{Siz1}$ and 4 μM PCNA at 4 °C. **b**, SDS–PAGE analysis of multiple turnover assays of SUMO modification of PCNA using *in vitro* reactions with coupled E1 (200 nM), E2 (100 nM), and E3 (50 nM) activities with 4 μM PCNA for the quantified data shown in Fig. 4d. **c**, SDS–PAGE analysis of multiple turnover assays of SUMO modification of PCNA using *in vitro* reactions with coupled E1 (200 nM), E2 (100 nM), and E3 (50 nM) activities with 4 μM PCNA and quantified. **d**, Representative non-reducing SDS–PAGE analysis of the single turnover *in vitro* assays of SUMO modification of PCNA shown in Fig. 4e. These assays utilized 5 nM of the $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488 thioester (or $E2_{Ubc9}$ mutant thioesters) in reactions with 50 nM of the indicated $E3_{Siz1}$ and a titration of PCNA. Shown are typical SDS–PAGE analyses from the 10 μM PCNA reactions. The data were used to extract the kinetic constants for the reactions shown as histograms and in Extended Data Table 5. For **a**, **c**, and **d** the quantified rate data show mean ± s.d. ($n = 3$ technical replicates). For gel source data, see Supplementary Fig. 1.



Extended Data Figure 6 | Shape complementarity between the E2$_{\text{Ubc9}}$–SUMO/E3 complex and PCNA. a, The current structure (colour) with the crystallographic packing of a lattice mate PCNA molecule (black). **b**, Non-reducing SDS–PAGE analysis of 2 min endpoint *in vitro* E2$_{\text{Ubc9}}$–SUMO thioester formation reactions with 0.05 µM E1, 0.4 µM of the indicated E2$_{\text{Ubc9}}$, and 22 µM SUMO (left) and the quantified E2–SUMO band (right). The quantified band intensity shows mean ± s.d. ($n = 3$ technical replicates). **c**, SDS–PAGE analysis of multiple turnover assays of SUMO modification of PCNA using *in vitro* reactions with coupled E1 (200 nM), E2 (100 nM), and E3 (50 nM) activities with 4 µM PCNA or without PCNA (diSUMO formation) shown quantified in Fig. 5c. **d**, Location of E2$_{\text{Ubc9}}$ and PCNA mutations that had no effect (red sticks) on activities observed for *in vitro* assays similar to those performed in Fig. 5c in relation to residues that did show effects (green sticks). **e**, The E2$^{\text{C87S}}_{\text{Ubc13}}$–Ub was aligned to E2$_{\text{Ubc9}}$ in the current structure and subsequently the Lys164/Glu165 loop from trimeric PCNA (pink) was aligned onto the Lys63/Glu64 loop from acceptor ubiquitin (PDB 2GMI, green). Within this conformation the E3$_{\text{Siz1}}$ PINIT domain (cyan) clashes with another protomer of the PCNA trimer (grey). For gel source data, see Supplementary Fig. 1.

Extended Data Table 1 | Summary of kinetic constants or inhibition constants for E2$_{\mathrm{Ubc9}}$–SUMO thioester and thioester mimetic association to E3$_{\mathrm{Siz1}}$

Thioester Mimetic		K_m (μM)	K_i (μM)
-	K127	1.07 ± 0.10	-
C93K	K127	0.96 ± 0.08	2.05 ± 0.19
A129K	K127	1.11 ± 0.07	4.33 ± 0.33
-	K164	0.70 ± 0.07	-
C93K	K164	0.72 ± 0.05	1.44 ± 0.13
A129K	K164	0.63 ± 0.05	3.15 ± 0.27

Calculated K_m for E2$_{\mathrm{Ubc9}}$–SUMO$^{\mathrm{D68R}}$–Alexa488 binding to E3$^{(167-465)}_{\mathrm{Siz1}}$ and calculated K_i values for competitive inhibition of this interaction by the indicated thioester mimetics using multiple turnover SUMO conjugation of PCNA at 30 °C with 100 nM–5 μM purified E2$_{\mathrm{Ubc9}}$–SUMO$^{\mathrm{D68R}}$–Alexa488 thioester with 1 nM E3 and 32 μM PCNA in the absence and presence of 2–20 μM of the indicated mimetic. Data show mean ± s.d. ($n = 3$ technical replicates).

Extended Data Table 2 | Summary of binding curve fits of fluorescent polarization data for Ubc9 binding Alexa488-labelled SUMO or SUMO^{D68R}

FP (Receptor Depletion Model)	SUMO-Alexa488	SUMO^{D68R}-Alexa488
Best Fit Values (+/- Std. Error)		
Limiting anisotropy free ligand	0.09 ± 2.77	-0.02 ± 0.51
Limiting anisotropy bound ligand	99.94 ± 1.67	100.00 ± 236.90
[Ligand]	= 50.00 nM	= 50.00 nM
K_d (nM)	24.6 ± 4.4	62,819 ± 169,826
95% Confidence Intervals		
Limiting anisotropy free ligand	-5.55 to 5.73	-1.07 to 1.02
Limiting anisotropy bound ligand	96.55 to 103.3	-381.9 to 581.9
K_d (nM)	15.57 to 33.62	0.0 to 408,332
Goodness of Fit		
Degrees of Freedom	33	33
R square	0.97	0.75
Absolute Sum of Squares	1253	188.4
Sy. x	6.16	2.39
Constraints		
[Ligand]	= 50.00 nM	= 50.00 nM
K_d	$K_d > 0.0$	$K_d > 0.0$
Number of Points		
Analyzed	36	36

Extended Data Table 3 | Summary of kinetic constants for multiple turnover experiments with purified E2$_{Ubc9}$–SUMO^{D68R} thioester

Siz1 Isoform	Additional SUMO Isoform	K$_{m}$ (μM)	k$_{cat}$ (s^{-1})	k$_{cat}$/ K$_{m}$ (M^{-1}/s^{-1})
167-465	-	1.07 ± 0.10	1.27 ± 0.05	1.18 x 10^{6} ± 0.12 x 10^{6}
167-449	-	1.23 ± 0.19	0.88 ± 0.05	7.15 x 10^{5} ± 0.12 x 10^{5}
167-449	Δ18Smt3ΔGGHIS	1.84 ± 0.31	0.86 ± 0.06	4.68 x 10^{5} ± 0.09 x 10^{5}
167-449	Δ18Smt3^{D68R}ΔGGHIS	1.08 ± 0.12	0.81 ± 0.03	7.49 x 10^{5} ± 0.09 x 10^{5}
167-449-Δ18Smt3ΔGGHIS	-	0.48 ± 0.15	0.64 ± 0.06	1.32 x 10^{6} ± 0.43 x 10^{6}
167-449-Δ18Smt3^{D68R}ΔGGHIS	-	4.40 ± 1.44	0.93 ± 0.18	2.11 x 10^{5} ± 0.08 x 10^{5}
167-508	-	2.34 ± 0.58	0.54 ± 0.06	2.31 x 10^{5} ± 0.06 x 10^{5}
167-508	Δ18Smt3ΔGGHIS	1.15 ± 0.14	0.89 ± 0.04	7.71 x 10^{5} ± 0.10 x 10^{5}
167-508	Δ18Smt3^{D68R}ΔGGHIS	1.46 ± 0.18	0.75 ± 0.04	5.13 x 10^{5} ± 0.07 x 10^{5}
167-508-Δ18Smt3ΔGGHIS	-	0.29 ± 0.07	0.71 ± 0.04	2.43 x 10^{6} ± 0.57 x 10^{6}
167-508-Δ18Smt3^{D68R}ΔGGHIS	-	2.07 ± 0.29	0.70 ± 0.04	3.38 x 10^{5} ± 0.05 x 10^{5}

Siz1 Isoform	Additional SUMO Isoform	K$_{m}$ (μM)	k$_{cat}$ (s^{-1})	k$_{cat}$/ K$_{m}$ (M^{-1}/s^{-1})
167-465	-	0.70 ± 0.07	4.26 ± 0.14	6.09 x 10^{6} ± 0.69 x 10^{6}
167-449	-	0.91 ± 0.12	2.52 ± 0.12	2.76 x 10^{6} ± 0.39 x 10^{6}
167-449	Δ18Smt3ΔGGHIS	1.44 ± 0.21	2.47 ± 0.15	1.71 x 10^{6} ± 0.27 x 10^{6}
167-449	Δ18Smt3^{D68R}ΔGGHIS	0.90 ± 0.07	2.45 ± 0.07	2.71 x 10^{6} ± 0.22 x 10^{6}
167-449-Δ18Smt3ΔGGHIS	-	0.23 ± 0.05	1.47 ± 0.08	6.48 x 10^{6} ± 1.58 x 10^{6}
167-449-Δ18Smt3^{D68R}ΔGGHIS	-	1.02 ± 0.19	1.29 ± 0.09	1.26 x 10^{6} ± 0.25 x 10^{6}
167-508	-	0.91 ± 0.12	1.12 ± 0.05	1.29 x 10^{6} ± 0.18 x 10^{6}
167-508	Δ18Smt3ΔGGHIS	0.62 ± 0.07	2.01 ± 0.07	3.22 x 10^{6} ± 0.39 x 10^{6}
167-508	Δ18Smt3^{D68R}ΔGGHIS	0.65 ± 0.10	1.47 ± 0.08	2.27 x 10^{6} ± 0.36 x 10^{6}
167-508-Δ18Smt3ΔGGHIS	-	0.14 ± 0.03	1.77 ± 0.08	13.09 x 10^{6} ± 2.81 x 10^{6}
167-508-Δ18Smt3^{D68R}ΔGGHIS	-	0.76 ± 0.10	1.34 ± 0.06	1.77 x 10^{6} ± 0.24 x 10^{6}

Top: catalytic constants for multiple turnover SUMO conjugation of PCNA to K127 at 30 °C with 100 nM–5 μM purified E2$_{Ubc9}$–SUMO^{D68R}–Alexa488 thioester with and without a 1.5-fold excess of an additional molecule of SUMO, 1 nM E3, and 32 μM PCNA. Data show mean ± s.d. ($n = 3$ technical replicates). Bottom: catalytic constants for multiple turnover SUMO conjugation of PCNA to K164 at 30 °C with 100 nM–5 μM purified E2$_{Ubc9}$–SUMO^{D68R}–Alexa488 thioester with and without a 1.5-fold excess of an additional molecule of SUMO, 1 nM E3, and 32 μM PCNA. Data show mean ± s.d. ($n = 3$ technical replicates).

Extended Data Table 4 | Data collection and refinement statistics

	$E2_{Ubc9}$-SUMO/$E3_{Siz1}$-SUMO/PCNA
Data collection	
Space group	C121
Cell dimensions	
a, *b*, *c* (Å)	93.42, 205.88, 142.50
α, β, γ (°)	90.00, 95.30, 90.00
Resolution (Å)	48.4-2.85 (2.95-2.85) *
R_{merge}	10.7 (55.6)
$I / \sigma I$	9.1 (1.82)
Completeness (%)	99.0 (98.0)
Redundancy	3.6 (2.9)
Refinement	
Resolution (Å)	47.3-2.85
No. reflections	61981
R_{work} / R_{free}	0.210/0.248
No. atoms	13744
Protein	13403
Ligand/ion	70
Water	271
B-factors	
Protein	58.3
Ligand/ion	64.5
Water	43.0
R.m.s. deviations	
Bond lengths (Å)	0.001
Bond angles (°)	0.42

A single crystal was used.
*Values in parentheses are for highest-resolution shell.

Extended Data Table 5 | Summary of kinetic constants for single turnover experiments with purified $E2_{Ubc9}$–$SUMO^{D68R}$ thioester

Ubc9- $Smt3^{D68R}$Alexa488 Isoform	$Siz1^{(167-465)}$ Isoform	PCNA Isoform	K_d (μM)	k_2 (s^{-1})	k_2/K_d (M^{-1}/s^{-1})	Kinetic Model	K_i (μM)
K154R (WT)	WT	WT	7.46 ± 1.15	$8.54x10^{-2} \pm 0.47x10^{-2}$	$1.14x10^{4} \pm 0.19x10^{4}$	S.I.*	1430 ± 690
K154R (WT)	WT	K164R	29.96 ± 5.85	$1.36x10^{-1} \pm 0.16x10^{-1}$	$4.54x10^{3} \pm 1.03x10^{3}$	S.I.*	190 ± 50
K154R (WT)	F268A	WT	39.15 ± 10.83	$4.99x10^{-2} \pm 0.72x10^{-2}$	$1.27x10^{3} \pm 0.40x10^{3}$	S.I.*	410 ± 140
Y87A/K154R	WT	WT	Below Detect.	Below Detect.	Below Detect.	-	-
Y87A/K154R	WT	K164R	33.66 ± 10.88	$1.17x10^{-4} \pm 0.23x10^{-4}$	3 ± 1	S.I.*	110 ± 40
D100A/K154R	WT	WT	53.92 ± 23.58	$6.38x10^{-2} \pm 1.52x10^{-2}$	$1.18x10^{3} \pm 0.59\ x10^{3}$	S.I.*	490 ± 280
S127A/K154R	WT	WT	36.96 ± 12.19	$3.22x10^{-3} \pm 0.50x10^{-3}$	90 ± 30	S.I.*	840 ± 450

Ubc9- $Smt3^{D68R}$Alexa488 Isoform	$Siz1^{(167-465)}$ Isoform	PCNA Isoform	K_d (μM)	k_2 (s^{-1})	k_2/K_d (M^{-1}/s^{-1})	Kinetic Model	K_i (μM)
K154R (WT)	WT	WT	7.47 ± 0.88	$1.39x10^{-1} \pm 0.04x10^{-1}$	$1.86x10^{4} \pm 0.23x10^{3}$	M.M.†	-
K154R (WT)	WT	K164R	-	-	-	-	-
K154R (WT)	F268A	WT	21.06 ± 5.26	$9.18x10^{-3} \pm 0.57x10^{-3}$	$4.36x10^{2} \pm 1.12x10^{2}$	M.M.†	-
Y87A/K154R	WT	WT	24.24 ± 5.65	$2.63x10^{-2} \pm 0.16x10^{-2}$	$1.08x10^{3} \pm 0.26x10^{3}$	M.M.†	-
Y87A/K154R	WT	K164R	-	-	-	-	-
D100A/K154R	WT	WT	54.40 ± 29.80	$1.03x10^{-2} \pm 0.18x10^{-2}$	190 ± 100	M.M.†	-
S127A/K154R	WT	WT	25.12 ± 6.93	$8.96x10^{-4} \pm 0.64x10^{-4}$	40 ± 10	M.M.†	-

Top: catalytic constants for single turnover SUMO conjugation of PCNA to K127 at 4 °C with 5 nM purified $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488 thioester, 50 nM E3, and 0.5–500 μM PCNA as indicated. Data show mean ± s.d. (*n* = 3 technical replicates). Bottom: catalytic constants for single turnover SUMO conjugation of PCNA to K164 at 4 °C with 5 nM purified $E2_{Ubc9}$–$SUMO^{D68R}$–Alexa488 thioester, 50 nM E3, and 0.5–500 μM PCNA as indicated. Data show mean ± s.d. (*n* = 3 technical replicates).

*Substrate inhibition.

†Michaelis–Menten.

doi:10.1038/nature19061

Dependence of the critical temperature in overdoped copper oxides on superfluid density

I. Božović[1,2], X. He[1,2], J. Wu[1] & A. T. Bollinger[1]

The physics of underdoped copper oxide superconductors, including the pseudogap, spin and charge ordering and their relation to superconductivity[1–3], is intensely debated. The overdoped copper oxides are perceived as simpler, with strongly correlated fermion physics evolving smoothly into the conventional Bardeen–Cooper–Schrieffer behaviour. Pioneering studies on a few overdoped samples[4–11] indicated that the superfluid density was much lower than expected, but this was attributed to pair-breaking, disorder and phase separation. Here we report the way in which the magnetic penetration depth and the phase stiffness depend on temperature and doping by investigating the entire overdoped side of the $La_{2-x}Sr_xCuO_4$ phase diagram. We measured the absolute values of the magnetic penetration depth and the phase stiffness to an accuracy of one per cent in thousands of samples; the large statistics reveal clear trends and intrinsic properties. The films are homogeneous; variations in the critical superconducting temperature within a film are very small (less than one kelvin). At every level of doping the phase stiffness decreases linearly with temperature. The dependence of the zero-temperature phase stiffness on the critical superconducting temperature is generally linear, but with an offset; however, close to the origin this dependence becomes parabolic. This scaling law is incompatible with the standard Bardeen–Cooper–Schrieffer description.

Using atomic-layer-by-layer molecular beam epitaxy (ALL-MBE; refs 12, 13), we synthesized single-crystal films of $La_{2-x}Sr_xCuO_4$ (LSCO), the simplest copper oxide that we can dope all the way to a non-superconducting metal state. As single-layer LSCO sustains high-temperature superconductivity (HTS) with a T_c value as high as is found in bulk samples[13], the physics is quasi-two-dimensional (2D) and we focus on the in-plane properties.

We used the mutual inductance technique[9,10,14–16] (Fig. 1a–c and Extended Data Figs 1 and 2), improved (see Methods) to resolve the absolute value of λ with a ±1% accuracy. For this, it is critical to accurately determine the superconducting layer thickness, which we achieve by engineering the samples at the atomic-layer level as illustrated in Fig. 1d. The sharpness of the peak in $\mathrm{Im}\,V_p(T)$, or equivalently in $\mathrm{Im}\,M(T)$, where V_p is the voltage on the pick-up coil and M is the mutual inductance, puts an upper bound of about 0.1 K on the spread in T_c (defined by the onset of the Meissner effect, that is, the expulsion of the magnetic field) in this $10 \times 10\,\mathrm{mm}^2$ film.

We have studied over 2,000 LSCO films in great detail (see Extended Data Figs 3–7). The film thickness was varied from $d = 0.66\,\mathrm{nm}$ (half of the unit cell height) to over 100 nm, and the composition was varied across the entire phase diagram. This was crucial for obtaining definitive conclusions—copper oxides are complex compounds, HTS has largely been a materials science endeavour, and good statistics are essential.

In Fig. 2 we show our key experimental data extracted directly from the measured inductance with dense coverage of the entire overdoped LSCO region. Figure 2a shows the doping dependence of $\lambda(T)$ for the hundred most homogeneous films, which are likely to represent the intrinsic LSCO properties. Figure 2b, c shows the 2D superfluid phase stiffness $\rho_s \equiv A/\lambda^2$, which is directly proportional to the 2D superfluid density $n_s^{2D} = \rho_s(4k_B m^*/\hbar^2)$. Here $A = \hbar^2 d/4\mu_0 k_B e^2 = 3.55 \times 10^{-12}\,\mathrm{m^2\,K^{-1}}$, k_B is the Boltzmann constant, m^* is the electron effective mass, $\hbar$ is the reduced Planck constant, $\mu_0 = 4\pi \times 10^{-7}\,\mathrm{N\,A^{-2}}$ is the vacuum permeability and e is the electron charge. The $\rho_s(T)$ dependence is essentially linear, even in the heavily overdoped LSCO films (Fig. 2c). A crossover to a T^2-dependence occurs below a very low sample-dependent temperature T^{**}. Figure 2d shows the dependence of T_c on ρ_{s0}. (The subscript 0 refers to the $T \to 0$ limit; this extrapolation is justified because our measurements extend down to $T = 300\,\mathrm{mK}$.) The $T_c(\rho_{s0})$ scaling in Fig. 2d is largely linear, $T_c = T_0 + \alpha\rho_{s0}$, with $T_0 = (7.0 \pm 0.1)\,\mathrm{K}$ and the proportionality coefficient $\alpha = 0.37 \pm 0.02$, except very close to the origin where the curve fits closely to $T_c = \gamma\sqrt{\rho_{s0}}$, where $\gamma = (4.2 \pm 0.5)\,\mathrm{K^{1/2}}$. The data are as accurate as depicted in the figure; the error bars are smaller than the marker size.

Our findings reinforce pioneering observations[4–7] of the diminishing ρ_{s0} in a few overdoped $Tl_2Ba_2CuO_{6-\delta}$ (Tl-2201) samples, and subsequently also in overdoped LSCO[8–11]. We confirm this result for

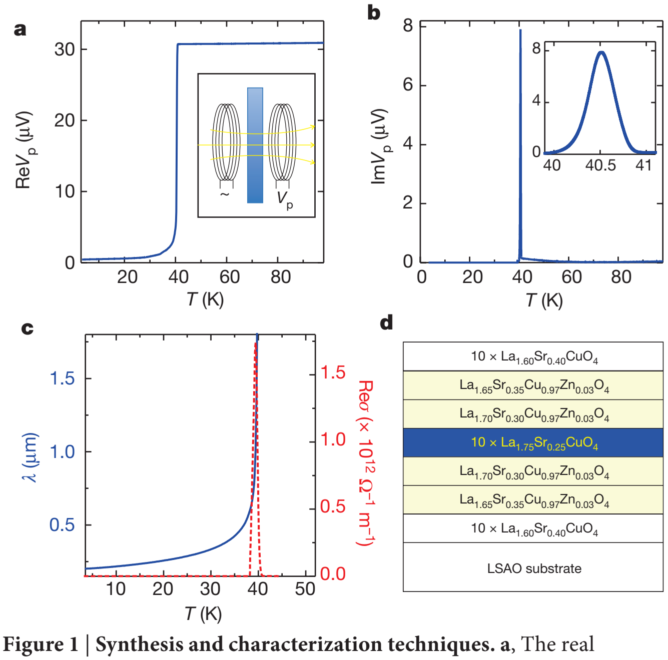


Figure 1 | Synthesis and characterization techniques. a, The real (in-phase) component of V_p, the voltage across the pickup coil (proportional to the mutual inductance), showing diamagnetic screening (the Meissner effect) when the film becomes superconducting. Inset, schematic of the experiment. **b**, The imaginary part of V_p shows that in this $10 \times 10\,\mathrm{mm}^2$ film, T_c is homogeneous to better than 0.1 K. Inset, the same, near T_c. **c**, The penetration depth, λ, and the real part of the complex a.c. ($\nu = 40\,\mathrm{kHz}$) conductivity, $\mathrm{Re}\sigma$, derived from the complex impedance. **d**, Schematic of a sample engineered for this study at the atomic-layer level. LSAO denotes the $LaSrAlO_4$ substrate.

[1]Brookhaven National Laboratory, Upton, New York 11973-5000, USA. [2]Applied Physics Department, Yale University, New Haven, Connecticut 06520, USA.

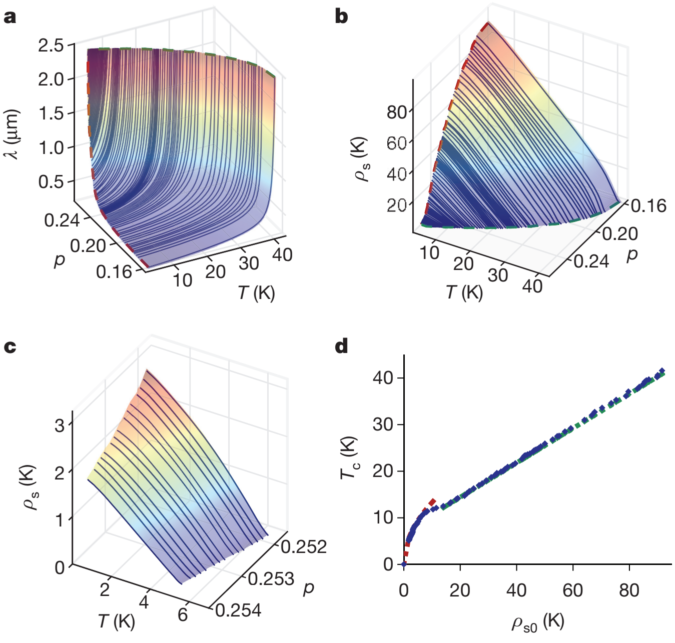


Figure 2 | The evolution of the superfluid with temperature and doping. **a**, The penetration depth $\lambda(T)$ measured in the 100 most homogeneous LSCO films synthesized by ALL-MBE. The red and green dashed lines, as well as the coloured shading, are visual aids. **b**, The corresponding phase stiffness $\rho_s(T)$. **c**, $\rho_s(T)$ for the most overdoped samples, measured down to $T = 0.3$ K. **d**, The dependence of T_c on $\rho_{s0} \equiv \rho_s(T \rightarrow 0)$. The experimental data are represented by the blue diamonds; the green dashed line is the fit to $T_c = T_0 + \alpha\rho_{s0}$ with $\alpha = 0.37 \pm 0.02$ for $\rho_{s0} > 15$ K and the red dashed line is the fit to $T_c = \gamma\sqrt{\rho_{s0}}$ with $\gamma = (4.2 \pm 0.5)$ K$^{1/2}$ for $\rho_{s0} < 12$ K. Source Data used to generate the lines in **b**–**d** are available online.

thousands of LSCO films and establish that ρ_{s0} decreases monotonously from the optimal doping until it vanishes and T_c drops to zero.

The disappearance of ρ_{s0} with overdoping was originally attributed to spontaneous electronic phase separation[4]. However, quantum oscillations observed in overdoped Tl-2201 with $T_c = 10$ K indicate electronic homogeneity on the 100 nm scale[17,18]. We show that LSCO films grown by ALL-MBE are also intrinsically homogeneous, at every doping level.

A nearly linear $\rho_s(T)$ dependence was observed[19] by a microwave technique in an overdoped Tl-2201 crystal with $T_c \approx 25$ K and very low T^{**} (2 K). We show that the same is true in LSCO films grown by ALL-MBE at every level of doping, with the slope independent of the carrier density p except very close to the edge of the dome-shaped $T_c(p)$ curve.

However, our central result, the $T_c(\rho_{s0})$ scaling law shown in Fig. 2d, differs qualitatively from these early inferences. The well-known Uemura's law, $T_c \propto n_{s0}/m^*$, inferred from muon spin rotation (μSR) measurements[4], refers to the underdoped side. On the overdoped side, the scarce μSR data were interpolated by a concave 'boomerang' shape[5–7]. In contrast, the $T_c(\rho_{s0})$ curve shown in Fig. 2d is markedly convex. The same discrepancy is found with Homes' law, $\rho_{s0} \propto \sigma_{dc}T_c$ (where σ_{dc} is the normal-state conductivity measured close to T_c), based on optics data[20]. According to Homes' law and our measured T_c and ρ_{s0} data, $\rho_{dc} \equiv 1/\sigma_{dc}$ should increase with doping and diverge as $1/\sqrt{\rho_{s0}}$ when $\rho_{s0} \rightarrow 0$, because under these assumptions $T_c \propto \sqrt{\rho_{s0}}$ (Extended Data Fig. 8). However, our measured values of ρ_{dc} decrease monotonously (essentially linearly) with doping. These discrepancies may originate from differences in the quality and homogeneity of the samples—copper oxides are complex materials and call for advanced synthesis techniques—and in much larger error bars for the ρ_{s0} values extracted from μSR and optics data.

A study of high-quality, heavily underdoped $YBa_2Cu_3O_{7-x}$ (YBCO) crystals by a microwave cavity-perturbation technique[21,22] reported $\rho_s(T)$ curves that are linear down to $T^{**} \approx 4$ K. The $T_c \propto \sqrt{\rho_{s0}}$ scaling was seen over a tiny range, $0.054 < p < 0.058$ (where p is the doping level), near the transition from superconductor to insulator and was attributed to quantum critical fluctuations. At higher doping levels, the $T_c(\rho_{s0})$ dependence became linear with a clear offset $T_0 \approx 4$–5 K. This is reminiscent of our results in Fig. 2, but in a different compound, indicating that this behaviour may be universal for the hole-doped copper oxides. Moreover, the fact that the results for underdoped YBCO are similar to those for overdoped LSCO is suggestive of underdoped/overdoped symmetry.

Our results, inferred from the raw data without any assumptions, challenge current thinking. In a clean Bardeen–Cooper–Schrieffer (BCS) superconductor, ρ_{s0} should be equal to the total particle density. (In fact, Leggett's theorem[23,24] asserts that this is true as long as superfluidity conforms to the two-fluid scenario, and more generally for any single-species system that is translation- and time-reversal-invariant.) Pair-breaking due to impurities and disorder can reduce ρ_{s0}, but we doubt that our data can be quantitatively explained using the standard 'dirty' d-wave BCS formalism, which includes the effects of pair breaking on impurities and other defects, because in that case the Glover–Ferrell–Tinkham sum rule implies that Homes' law should apply[20,25], contrary to what we observe. Qualitatively, for ρ_{s0} to vanish because of disorder the sample must get extremely dirty and a superconductor-to-insulator transition would be observed, while in fact Fig. 3a shows LSCO becoming more metallic.

The $\rho_s(T)$ dependence is expected to be linear only in very clean d-wave BCS superconductors; disorder and pair-breaking turn this relation quadratic[21,22,26] below T^{**}. In Fig. 2c, $T^{**} \lesssim 2$ K, comparable to the high-quality YBCO crystals[21,22] and corresponding within this formalism to a mean-free-path $l_0 \gtrsim 4$ μm and thus to the 'ultra-clean' limit. In Fig. 3b, we illustrate the effect of (deliberately added) disorder by comparing two samples with the same T_c (38 K), one slightly overdoped ($p = 0.19$) and clean, and the other optimally doped ($p = 0.16$) but with 0.5% of the Cu replaced by Zn, a known pair-breaker. Although this small amount of disorder does not affect T_c greatly ($\Delta T_c \approx -3$ K), it has a dramatic effect on the shape of the $\rho_s(T)$ curve, which becomes parabolic below 20–25 K. We are therefore able to detect even a small amount of disorder and pair-breaking.

The results we highlight here—the linear $\rho_s(T)$ dependence and the overall $T_c(\rho_{s0})$ scaling—are robust and not sample-dependent, and are hence likely to be intrinsic. (On the contrary, T^{**} and ρ_{dc} are affected by disorder and other extrinsic factors; both can be modified by annealing the film in a vacuum, oxygen, or ozone, thus modifying the concentration of oxygen vacancies, without a substantial effect on T_c.) The extremely dirty BCS picture is also inconsistent with observations of quantum oscillations and other experiments[17,18,27].

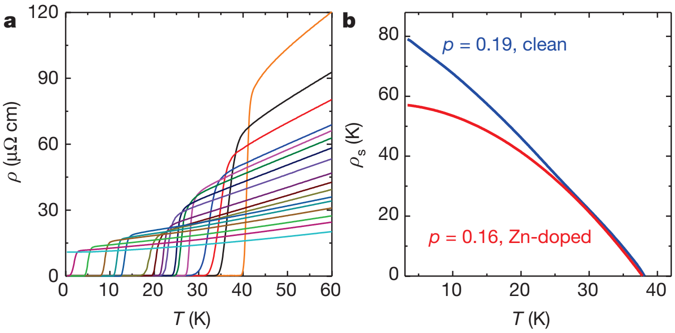


Figure 3 | Overdoped LSCO films synthesized by ALL-MBE are clean superconductors. **a**, Resistivity in LSCO films with different doping levels (see Methods), top to bottom: $p = 0.172$, 0.205, 0.211, 0.217, 0.220, 0.224, 0.228, 0.230, 0.233, 0.237, 0.242, 0.245, 0.248, 0.251, 0.254, 0.258, 0.295. **b**, The dependence of the shape of $\rho_s(T)$ on disorder. The red line represents an LSCO film that is optimally doped ($p = 0.16$) but with 0.5% of the Cu substituted by Zn. The blue line indicates a clean LSCO film that is slightly overdoped ($p = 0.19$) to have the matching T_c (approximately 38 K). In the clean film the $\rho_s(T)$ dependence is essentially linear, while in the dirty film pair-breaking makes this relation parabolic below about 25 K.

Homes' law can be derived by a different line of reasoning[25], assuming the 'Planckian' dissipation, with the scattering time $\tau = \hbar/k_B T$ as strong as is allowed by the uncertainty relation—which connects it naturally to another mysterious feature of HTS copper oxides, the linear temperature dependence of resistivity. The latter is probably related to the linear $\rho_{s0}(T)$ and $T_c(\rho_{s0})$ dependences we observe; as T_c and ρ_{s0} decrease with overdoping, the T-linear term in resistivity decreases as well[28], and this may be the reason for the deviation from the Homes relation.

Another important inference is that T_c seems to be principally controlled by the superfluid density. It seems unlikely that the doping level is the primary factor that controls both T_c and ρ_{s0}, because T_c strongly increases in LSCO under hydrostatic pressure or compressive epitaxial strain, which increases ρ_{s0} while keeping p constant[29]. The underdoped/overdoped symmetry, even if only approximate, is another strong indication that ρ_{s0}, rather than p, controls T_c. If T_c is in fact essentially determined by the kinematics, this points to local pairing rather than to BCS physics. This notion is further supported by extensive study of the magnetoresistance in thousands of LSCO samples (I.B., J.W., A.T.B. and X.H., unpublished results), which together with the results presented here leads to the conclusion that the pair size is always smaller than their separation. However, this premise alone does not resolve the paradoxes, because Leggett's theorem[23,24] remains valid for arbitrary interaction strengths. The fact that T_c and ρ_{s0} are comparable points to massive phase fluctuations[11,30], which calls any mean-field description into question. Our experimental findings therefore challenge the existing theories; the accurate scaling reported here may be a benchmark test.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 3 January; accepted 16 June 2016.

1. Lee, P. A., Nagaosa, N. & Wen, X. G. Doping a Mott insulator: physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78,** 17–85 (2006).
2. Zaanen, J. *et al.* Towards a complete theory of high T_c. *Nat. Phys.* **2,** 138–143 (2006).
3. Keimer, B., Kivelson, S. A., Norman, M. R., Uchida, S. & Zaanen, J. From quantum matter to high-temperature superconductivity in copper oxides. *Nature* **518,** 179–186 (2015).
4. Uemura, Y. J. *et al.* Universal correlations between T_c and n_s/m^* (carrier density over effective mass) in high-T_c cuprate superconductors. *Phys. Rev. Lett.* **62,** 2317–2320 (1989).
5. Uemura, Y. J. *et al.* Magnetic-field penetration depth in $Tl_2Ba_2CuO_{6+\delta}$ in the overdoped regime. *Nature* **364,** 605–607 (1993).
6. Niedermayer, C. *et al.* Muon spin rotation study of the correlation between T_c and n_s/m^* in overdoped $Tl_2Ba_2CuO_{6+\delta}$. *Phys. Rev. Lett.* **71,** 1764–1767 (1993).
7. Bernhard, C. *et al.* Magnetic penetration depth and condensate density of cuprate high-T_c superconductors determined by muon-spin-rotation experiments. *Phys. Rev. B* **52,** 10488–10498 (1995).
8. Panagopoulos, C. *et al.* Superfluid response in monolayer high-T_c cuprates. *Phys. Rev. B* **67,** 220502 (2003).
9. Locquet, J. P. *et al.* Variation of the in-plane penetration depth λ_{ab} as a function of doping in $La_{2-x}Sr_xCuO_{4+\delta}$ thin films on $SrTiO_3$: implications for the overdoped state. *Phys. Rev. B* **54,** 7481–7488 (1996).
10. Lemberger, T. R. *et al.* Superconductor-to-metal quantum phase transition in overdoped $La_{2-x}Sr_xCuO_4$. *Phys. Rev. B* **83,** 140507 (2011).
11. Rourke, P. *et al.* Phase fluctuating superconductivity in overdoped $La_{2-x}Sr_xCuO_4$. *Nat. Phys.* **7,** 455–458 (2011).
12. Bozovic, I. Atomic-layer engineering of superconducting oxides: yesterday, today, tomorrow. *IEEE Trans. Appl. Supercond.* **11,** 2686–2695 (2001).
13. Logvenov, G., Gozar, A. & Bozovic, I. High-temperature superconductivity in a single copper-oxygen plane. *Science* **326,** 699–702 (2009).
14. Hebard, A. F. & Fiory, A. T. Evidence for the Kosterlitz–Thouless transition in thin superconducting aluminum films. *Phys. Rev. Lett.* **44,** 291–294 (1980).
15. Claassen, J. H., Reeves, M. E. & Soulen, R. J. Jr. A contactless method for measurement of the critical current density and critical temperature of superconducting rings. *Rev. Sci. Instrum.* **62,** 996–1004 (1991).
16. Clem, J. R. & Coffey, M. W. Vortex dynamics in a type-II superconducting film and complex linear-response functions. *Phys. Rev. B* **46,** 14662–14674 (1992).
17. Vignolle, B. *et al.* Quantum oscillations in an overdoped high-T_c superconductor. *Nature* **455,** 952–955 (2008).
18. Bangura, A. F. *et al.* Fermi surface and electronic homogeneity of the overdoped cuprate superconductor $Tl_2Ba_2CuO_{6+\delta}$ as revealed by quantum oscillations. *Phys. Rev. B* **82,** 140501(R) (2010).
19. Deepwell, D. *et al.* Microwave conductivity and superfluid density in strongly overdoped $Tl_2Ba_2CuO_{6+\delta}$. *Phys. Rev. B* **88,** 214509 (2013).
20. Homes, C. C. *et al.* A universal scaling relation in high-temperature superconductors. *Nature* **430,** 539–541 (2004).
21. Hosseini, A. *et al.* Microwave spectroscopy of thermally excited quasiparticles in $YBa_2Cu_3O_{6.99}$. *Phys. Rev. B* **60,** 1349–1359 (1999).
22. Broun, D. M. *et al.* Superfluid density in a highly underdoped $YBa_2Cu_3O_{6+y}$ superconductor. *Phys. Rev. Lett.* **99,** 237003 (2007).
23. Leggett, A. *Quantum Liquids* Ch. 3.3–3.4 (Oxford Univ. Press, 2006).
24. Leggett, A. On the superfluid fraction of an arbitrary many-body system at $T = 0$. *J. Stat. Phys.* **93,** 927–941 (1998).
25. Zaanen, J. Superconductivity: why the temperature is high. *Nature* **430,** 512–513 (2004).
26. Hirschfeld, P. J. & Goldenfeld, N. Effect of strong scattering on the low-temperature penetration depth of a *d*-wave superconductor. *Phys. Rev. B* **48,** 4219–4222 (1993).
27. Alldredge, J. W. *et al.* Evolution of the electronic excitation spectrum with strongly diminishing hole density in superconducting $Bi_2Sr_2CaCu_2O_{8+\delta}$. *Nat. Phys.* **4,** 319–326 (2008).
28. Cooper, R. A. *et al.* Anomalous criticality in the electrical resistivity of $La_{2-x}Sr_xCuO_4$. *Science* **323,** 603–607 (2009).
29. Bozovic, I., Logvenov, G., Belca, I., Narimbetov, B. & Sveklo, I. Epitaxial strain and superconductivity in $La_{2-x}Sr_xCuO_4$ thin films. *Phys. Rev. Lett.* **89,** 107001 (2002).
30. Emery, V. & Kivelson, S. A. Importance of phase fluctuations in superconductors with small superfluid density. *Nature* **374,** 434–437 (1994).

Acknowledgements A. Gozar, J. Zhang and J. Yoon contributed to developing the characterization techniques during the early stages of this work. R. Sundling developed the software for the inversion of the inductance data. We also benefited from the electrolyte-gating experiments and X-ray diffraction studies by X. Leng, and from numerical simulations by N. Božović. The research was done at BNL and was supported by the US Department of Energy, Basic Energy Sciences, Materials Sciences and Engineering Division. X.H. is supported by the Gordon and Betty Moore Foundation's EPiQS Initiative through grant GBMF4410. I.B. acknowledges discussions with J. Zaanen, G. Deutscher, A. Leggett, P. Littlewood, C.-B. Eom, J. Mannhart, P. Coleman, R. Prozorov, D. van der Marel, A. McKenzie, V. Kogan, P. Armitage, J.-M. Triscone, P. Canfield, A. Chubukov, B. Halperin, P. Kim, T. Lemberger, M. V. Sadovskii, D. Pavuna, Z. Radović and M. Vanević.

Author Contributions I.B. conceived the project, synthesized the films using ALL-MBE, measured the inductance, analysed the data and wrote the text. X.H. synthesized the films, performed AFM imaging and measured the inductance. A.T.B. fabricated the devices by lithography and performed the inductance measurements in the ^{3}He system. J.W. performed the transport measurements.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to I.B. (bozovic@bnl.gov).

METHODS

Atomic-layer-by-layer molecular beam epitaxy (ALL-MBE) synthesis. In most HTS experiments so far, the main sources of uncertainty were the samples themselves. In complex materials such as copper oxides, some level of inhomogeneity is present in most samples due to extrinsic factors. Most HTS 'single crystals' in fact contain stacking faults and intergrowths of other copper oxide phases and polytypes. Moreover, oxygen is volatile in copper oxides, and hence bulk crystals are prone to gradients in the density of the oxygen vacancies or interstitials. Irregular geometries of crystals and/or contacts cause some uncertainty in transport property measurements. In principle, one can alleviate the above problems by working with very thin single-crystal films; however, most HTS films are granular and contain both secondary phase precipitates and pinholes. This calls for some advanced materials science—as well as for large sample sets with sufficient statistics to clearly discern intrinsic behaviour.

For film synthesis we use a custom ALL-MBE system[12]. It is equipped with 16 metal sources (either K cells or rod-fed electron beam sources), a pure ozone source and a 16-channel real-time rate monitoring system based on atomic absorption spectroscopy. It also contains a dual-deflection reflection high-energy electron diffraction (RHEED) system that can monitor 20 samples in parallel, and a time-of-flight ion scattering and recoil spectroscopy (TOF-ISARS) system for chemical analysis of the film surface. These advanced surface science tools provide real-time information about the morphology of the film surface, the chemical composition and the crystal structure, and are also quintessential to grow atomically smooth and perfect films[31–33].

Using this system, we have performed over 2,500 LSCO film growth experiments. Each film was characterized in real time by RHEED and *ex situ* by atomic force microscopy (AFM) and magnetic susceptibility measurements. RHEED oscillations provide a digital count of the atomic layers and real-time control of the film quality. Selected films were also characterized *in situ* by TOF-ISARS and *ex situ* by X-ray diffraction (XRD), transport measurements and Rutherford backscattering. Further characterization was undertaken using atomic-resolution scanning transmission electron microscopy and electron energy-loss spectroscopy, resonant elastic and inelastic synchrotron X-ray scattering, synchrotron-based X-ray phase-retrieval techniques such as coherent Bragg rod analysis, muon spin rotation, ultrafast electron diffraction, ultrafast optical and THz pump-probe techniques and so on.

Using ALL-MBE, we synthesize single-crystal thin films of LSCO that are atomically smooth, without any secondary phase precipitates or pinholes. The films can be made ultrathin, down to a single unit cell thick[31–33]; this is advantageous for transmission measurements, because a relatively large transmittance helps to minimize the effect of any small-area defects such as pinholes, which can arise from imperfections in substrate polishing, for example. Much thicker films are also grown for reflectance measurements.

To alleviate the problem of oxygen non-uniformity, we have performed over a thousand experiments that involved annealing in ozone, oxygen, or a vacuum, spanning 13 orders of magnitude in pressure from 10^{-8} torr to 200 atm. Before and after each annealing step the films were characterized by AFM, transport and XRD measurements. On the basis of these extensive studies, we have developed recipes that involve multiple annealing steps at different temperatures and pressures, yielding the most homogeneous films with the sharpest superconducting transitions.

Atomic-layer engineering. Uncertainty in the film thickness is another key problem that limits the accuracy of some measurements—of transmittance or the critical current density, for example. ALL-MBE solves this problem by providing digital control over the film thickness[31–33]; we count atomic monolayers, while we determine the lattice constant with crystallographic accuracy by X-ray diffraction. However, this still leaves some uncertainty about the actual thickness of the superfluid, because we have found that typically a couple of layers next to the substrate are modified structurally and chemically, and are not superconducting. The same is generally true for a couple of layers near the free film surface once the film has been exposed to contamination from the atmosphere. To eliminate this problem we resort to atomic-layer engineering.

An example is illustrated in Fig. 1d. The active (superconducting) part of the sample under study is an exactly 5-unit-cell- (5-UC-) thick HTS layer. It is protected on both sides by a metallic ($M = \mathrm{La}_{1.60}\mathrm{Sr}_{0.40}\mathrm{CuO}_4$) buffer and a cover layer of fixed thickness. However, without further sample engineering, there would be some hole depletion from M and accumulation in the nearest HTS layers[31,33]. To minimize this interfacial effect, we 'sculpt' the charge profile near the interfaces by graded Sr doping in the four relevant layers. Moreover, we also dope these transition layers by substituting 3% of the Cu with Zn; this is known[13] to suppress T_c by a factor of two and n_s even more dramatically, by four–five times. In this way, we quench any potential residual superconductivity and eliminate any interface contributions. One is then left with exactly 5 UC of superconducting material—10 HTS CuO_2 planes. We have double-checked the validity of this approach by synthesizing a series of heterostructures in which we kept the composition of the constituent materials unaltered, but varied the thickness of the active HTS layers and verified that the sheet superfluid density scaled linearly with the number of HTS CuO_2 planes (see Extended Data Fig. 1).

Note that this is just one example; indeed, we have synthesized a large number of samples, varying the layering scheme, the composition and thickness of individual layers, as well as the synthesis and post-annealing conditions. The results presented in this Letter are valid generally and not restricted to any of these choices.

Penetration depth measurements. The λ is typically measured using μSR[4–8,34], microwave resonance (cavity-perturbation)[20–22,35–37], or mutual inductance[9,10,15–17,38–54] techniques. Because of the cost and duration, μSR experiments are usually restricted to just a few compositions and temperatures. The microwave technique offers an unsurpassed relative accuracy but the absolute accuracy is limited by the uncertainty in geometric factors. For these reasons, and because it is best suited for thin-film studies, we have chosen the mutual inductance method, which has been pioneered by several groups[10,14,15] (Fig 1a, b). A detailed theoretical treatment[16] of a superconducting film of thickness d and infinite radius, characterized by the ac conductivity $\sigma(\omega) = \sigma_1(\omega) - i\sigma_2(\omega)$, where ω is the measurement (angular) frequency and i the imaginary unit, placed between two coils of radii R_1 and R_2, respectively, parallel to one another, and separated by a distance D, provided the following expression for the mutual inductance:

$$\begin{aligned}\hat{M} &= \mathrm{Re}M + i\mathrm{Im}M \\ &= \mu_0 \pi R_1 R_2 \frac{\int_0^\infty \mathrm{d}\boldsymbol{q}[\exp(-\boldsymbol{q}D)J_1(\boldsymbol{q}R_1)J_1(\boldsymbol{q}R_2)]}{\cosh(Qd) + [(Q^2 + \boldsymbol{q}^2)/2\boldsymbol{q}Q]\sinh(Qd)}\end{aligned} \qquad (1)$$

where $\mu_0 = 4\pi \times 10^{-7}$ F m^{-1}, $\boldsymbol{q}$ is the wave vector, $J_1(x)$ is the first-order Bessel function, $Q^2 = \boldsymbol{q}^2 + (1/\lambda^2) - i\mu_0\omega\sigma_1$ and $\sigma_2 = 1/\mu_0\omega\lambda^2$. This is readily generalized to the case of two solenoids with N_1 and N_2 turns, respectively, by the summation over each pair of coils, one from each side. Once the values of ReM and ImM are measured experimentally, using equation (1) one can determine the values of λ and σ_1.

In practice, this procedure is subject to some uncertainties because of imprecision in the coil geometry, run-to-run variations in the film position, the finite size of the film that allows some flux to 'leak' around it, other parasitic coupling between the two solenoids and inaccuracy in the film thickness. In addition, the measurements are done down to some finite temperature, most frequently $T = 4.2$ K, and the limiting value λ_0 is obtained by extrapolation. These uncertainties have been analysed in detail in the literature, and various solutions were suggested to reduce them. In what follows, we explain briefly the improvements that we made in the measuring apparatus and technique to tighten the error bars.

Parasitic field coupling around the film and through the electronics. We use inductance coils with a large number of turns (300–1,500) but very small inner radius (250 μm), much smaller than the film size (10×10 mm^2), so that field leakage around the film is minimal (<0.3%) to begin with. We nevertheless subtract it accurately based on measurements of films of Nb, Pb and Al that are thick enough for transmittance to be negligible, and deposited on identical substrates. Alternatively, we deposit a thick Al overlayer on top of the HTS film and switch the diamagnetic screening in Al (at $T = 0.3$ K) between total (zero transmittance) and essentially none (total transmittance) by applying a small d.c. magnetic field (typically 100 G) that drives Al normal but does not affect the HTS film. As a consistency check, we verified that the corrected N_{s0} scales linearly with the film thickness, while being independent of frequency. (In contrast, the leakage contribution varies with frequency ν, linearly for $10\,\mathrm{kHz} < \nu < 100\,\mathrm{kHz}$, and thus can be clearly identified by repeating measurements on the same film at several frequencies.) To minimize eddy currents, we built the sample holder out of a single block of sapphire crystal. The coils are fixed rigidly and the sample is spring-loaded so that the film surface always gets to exactly the same position. Overall, we have achieved reproducibility and precision of ±0.3%, or better.

Uncertainty in the coil geometry. Our measured mutual inductance $M(T)$ differs a little (typically by a couple of per cent) from the calculated value due to some deviations in the geometry of the actual coil from its idealized mathematical model. To account for this (multiplicative) factor, we normalize the measured $M(T)$ by its value M_{high} just above T_c, or equivalently by M_{sub} of a pristine substrate (the latter is essentially independent of temperature and equal to M_{high}).

Field penetration through secondary phase precipitates, scratches and pinholes. This is particularly dangerous for thicker films where the intrinsic transmission is small. For this reason, we synthesize very thin films by ALL-MBE, down to 0.5 UC thick. Also, we have a very large statistics, so we can easily recognize extrinsic behaviour and factors.

Uncertainty in the film thickness. This is probably the single largest source of error[40,45] in the measurements of λ by inductance techniques reported so far. We have minimized this using digital synthesis as described in the atomic-layer engineering section above, and illustrated in Fig. 1d. We also compared films of the same composition and T_c but of different thickness, as illustrated in Extended Data Fig. 1, and verified that the results are consistent. This constitutes the best confirmation that we can achieve at present. Regrettably, the technique described in Fig. 1d would not work on the underdoped side, but that does not affect our main goal here, which was to study the evolution of LSCO physics from the overdoped metal to the optimal doping level.

Extrapolation to $T \to 0$ of measurements done only down to $T = 4$ K. Although this has been done routinely in the literature, it may introduce error, especially for films with very low T_c—say, below about 10 K. As we are interested in behaviour near the quantum critical points where T_c vanishes, we have built a ^{3}He-based set-up that extends the temperature range down to 300 mK. In the same set-up, it is also possible to apply a d.c. magnetic field up to 9 T. To ensure the accuracy in temperature reading, we measure on both cooling and heating, at a very low rate of 0.1–0.2 K min^{-1}, in a set-up where we achieve[55] a thermal stability better than 1 mK.

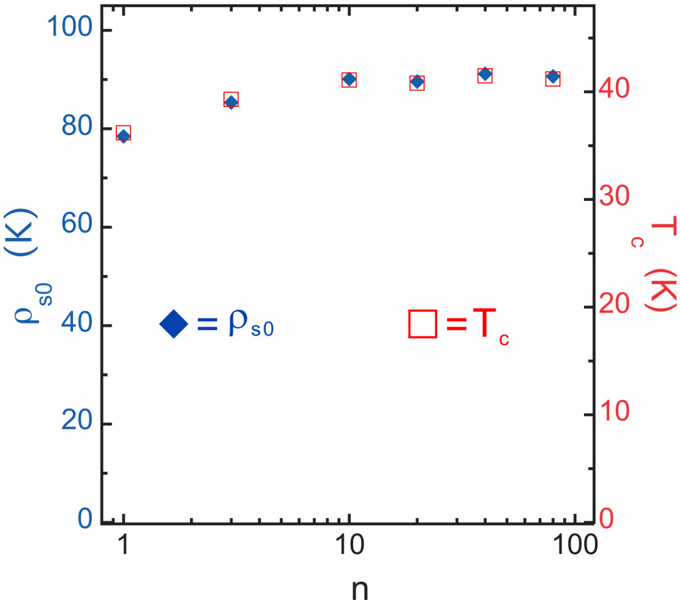
We measure $\lambda(T)$ with a reproducibility better than ±0.3%, as illustrated for a Nb film in Extended Data Fig. 2. Altogether, the accuracy in the absolute value of λ is better than ±1%. Moreover, some of our conclusions derive from the temperature dependence of λ, which is measured with a relative accuracy better than ±0.1%. The same is true of ρ_s, as $\rho_s = A/\lambda^2$, and $A = \hbar^2 d/4\mu_0 k_B e^2 = 3.55 \times 10^{-12}$ m^2 K^{-1}.

On selected films, we also compared our results with in-house high-frequency (0.5–50 MHz) inductance measurements in the reflectance geometry[52,53], as well as with the results of terahertz pump-probe[56] and μSR experiments[34] performed by our collaborators, and found good agreement.

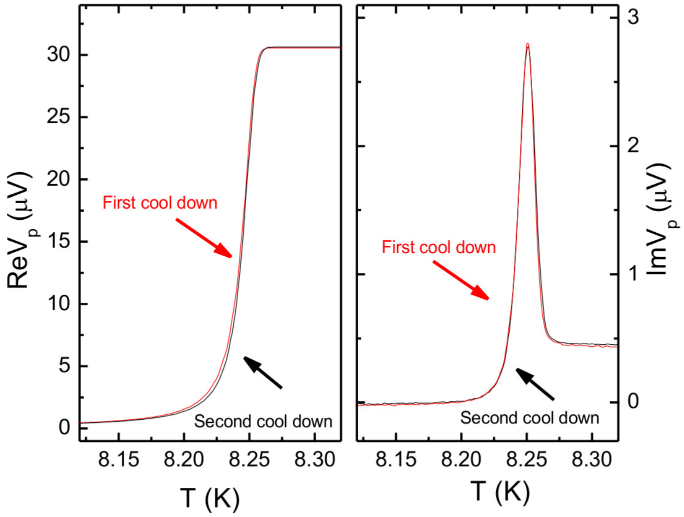
So far, we have performed inductance measurements on more than 2,500 LSCO films, a number of which were measured dozens of times and on multiple (a total of ten so far) set-ups. Mining this large database allows one to identify clear statistical trends and uncover intrinsic behaviour. In this Letter, we focus on the films with the sharpest transitions; in the best ones, near T_c we see $\sigma_1(\omega, T)$ rising exponentially on the scale of 0.1–0.2 K. This puts an upper bound on any inhomogeneity in T_c, as the transition width comes largely from thermal fluctuations. Clearly these samples are very homogeneous and hence they can be presumed to display intrinsic properties and behaviour.

The mobile carrier density. Note that the values of p quoted here, and in the literature, are only approximately equal to the mobile carrier (hole) density. A prevailing convention in the field is to infer p from the measured T_c, assuming a parabolic $T_c(p)$ relation. This label p should roughly indicate the doping state in a particular sample. To make it easier for the reader to compare our data with various phase diagrams in the literature, in Figs 2 and 3 we provide the nominal p values inferred by assuming that in LSCO $T_c = A(p - p_{c1})(p_{c2} - p)$, with $p_{c1} = 0.06$, $p_{c2} = 0.26$ and $A = 4.15 \times 10^3$ K, so that $p \equiv 0.16 + (0.01 - 2.4 \times 10^{-4} T_c)^{1/2}$. However, we stress that all of our conclusions are in fact only based on the quantities (T_c, λ, ρ_s and so on) that we measure directly and accurately.

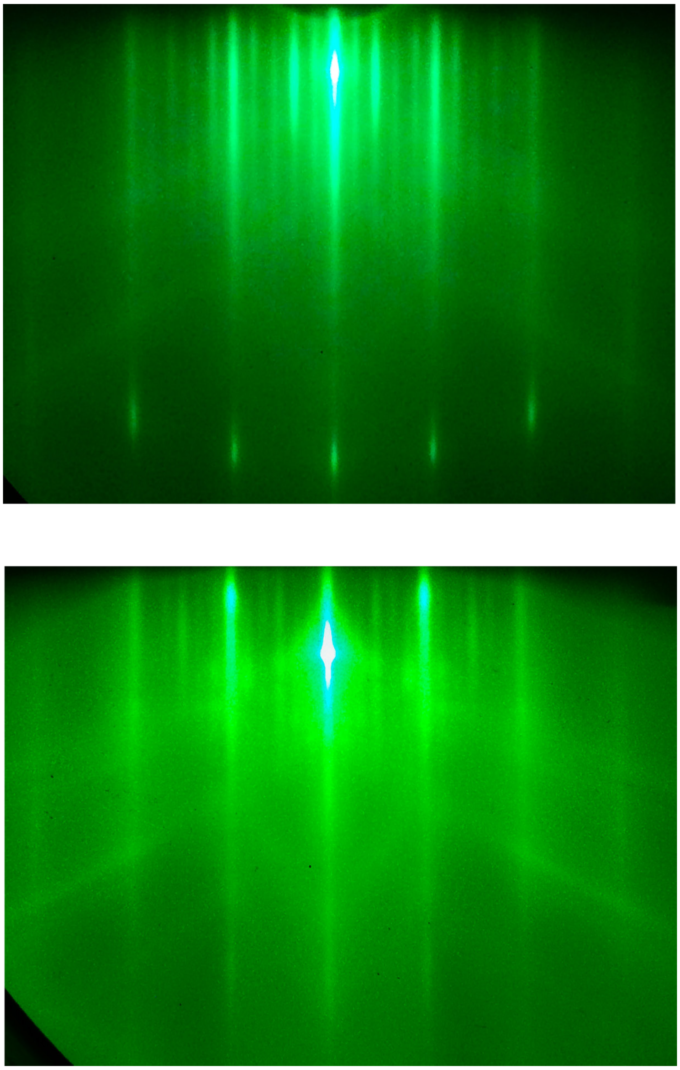
31. Gozar, A. *et al.* Interface superconductivity between a metal and a Mott insulator. *Nature* **455,** 782–785 (2008).
32. Bollinger, A. T. *et al.* Superconductor–insulator transition in $La_{2-x}Sr_xCuO_4$ at the pair quantum resistance. *Nature* **472,** 458–460 (2011).
33. Wu, J. *et al.* Anomalous independence of interface superconductivity on carrier density. *Nat. Mater.* **12,** 877–881 (2013).
34. Morenzoni, E. *et al.* The Meissner effect in a strongly underdoped cuprate above its critical temperature. *Nat. Commun.* **2,** 272 (2010).
35. Bonn, D. A. & Hardy, W. N. in *Physical Properties of High Temperature Superconductors* Vol. V (ed. Ginsberg, D. M.) 7–98 (World Scientific, 1996).
36. Prozorov, R. & Giannetta, R. W. Magnetic penetration depth in unconventional superconductors. *Supercond. Sci. Technol.* **19,** R41–R67 (2006).
37. Prozorov, R. & Kogan, V. G. London penetration depth in iron-based superconductors. *Rep. Prog. Phys.* **74,** 124505 (2011).
38. Lee, J. Y. & Lemberger, T. R. Penetration depth $\lambda(T)$ of $YBa_2Cu_3O_{7-\delta}$ films determined from the kinetic inductance. *Appl. Phys. Lett.* **62,** 2419–2421 (1993).
39. Pippard, A. B. Magnetic penetration depth through a superconducting film. *Supercond. Sci. Technol.* **7,** 696–699 (1994).
40. Turneaure, S. J., Ulm, E. R. & Lemberger, T. R. Numerical modeling of a two-coil apparatus for measuring the magnetic penetration depth in superconducting films and arrays. *J. Appl. Phys.* **79,** 4221–4227 (1996).
41. Fuchs, A., Prusseit, W., Berberich, P. & Kinder, H. High-precision penetration-depth measurement of $YBa_2Cu_3O_{7-\delta}$ as a function of oxygen content. *Phys. Rev. B* **53,** R14745–R14748 (1996).
42. Gilchrist, J. & Brandt, E. H. Screening effect of Ohmic and superconducting planar thin films. *Phys. Rev. B* **54,** 3530–3544 (1996).
43. Lee, J. Y., Kim, Y. H., Hahn, T.-S. & Choi, S. S. Determining the absolute value of penetration depth of large area films. *Appl. Phys. Lett.* **69,** 1637–1639 (1996).
44. Claassen, J. H., Wilson, M. L., Byers, J. M. & Adrian, S. Optimizing the two-coil mutual inductance measurement of the superconducting penetration depth in thin films. *J. Appl. Phys.* **82,** 3028–3034 (1997).
45. Turneaure, S. J., Pesetski, A. A. & Lemberger, T. R. Numerical modeling and experimental considerations for a two-coil apparatus to measure the complex conductivity of superconducting films. *J. Appl. Phys.* **83,** 4334–4343 (1998).
46. Paget, K. M. *et al.* Magnetic penetration depth in superconducting $La_{2-x}Sr_xCuO_4$ films. *Phys. Rev. B* **59,** 641–646 (1999).
47. Wang, R. F., Zhao, S. P., Chen, G. H. & Yang, Q. S. Absolute measurement of penetration depth in a superconducting film by the two-coil technique. *Appl. Phys. Lett.* **75,** 3865–3867 (1999).
48. Coffey, M. W. Analyzing mutual inductance measurements to determine the London penetration depth. *J. Appl. Phys.* **87,** 4344–4351 (2000).
49. Coffey, M. W. Mutual inductance of superconducting thin films. *J. Appl. Phys.* **89,** 5570–5577 (2001).
50. Rüfenacht, A., Locquet, J. P., Fompeyrine, J., Caimi, D. & Martinoli, P. Electrostatic modulation of the superfluid density in an ultrathin $La_{2-x}Sr_xCuO_4$ film. *Phys. Rev. Lett.* **96,** 227002 (2006).
51. Lemberger, T. R., Hetel, I., Tsukada, A. & Naito, M. Anomalously sharp superconducting transitions in overdoped $La_{2-x}Sr_xCuO_4$ films. *Phys. Rev. B* **82,** 214513 (2010).
52. Gauzzi, A. *et al.* Very high resolution measurement of the penetration depth of superconductors by a novel single-coil inductance technique. *Rev. Sci. Instrum.* **71,** 2147–2153 (2000).
53. Gasparov, V. A. & Bozovic, I. Magnetic field and temperature dependence of complex conductance of ultrathin $La_{1.55}Sr_{0.45}CuO_4/La_2CuO_4$ films. *Phys. Rev. B* **86,** 094523 (2012).
54. Došlić, M., Pelc, D. & Požek, M. Contactless measurement of nonlinear conductivity in the radio-frequency range. *Rev. Sci. Instrum.* **85,** 073905 (2014).
55. Dubuis, G., He, X. & Božović, I. Ultra-thermal-stabilization of a closed cycle cryocooler. *Rev. Sci. Instrum.* **85,** 103902 (2014).
56. Bilbro, L. S. *et al.* Temporal correlations of superconductivity above the transition temperature in $La_{2-x}Sr_xCuO_4$ probed by terahertz spectroscopy. *Nat. Phys.* **7,** 298–302 (2011).
57. Donnelly, R. J. in *Physics Vade Mecum* (ed. Anderson, H. L.) 121, Table E (AIP New York 1981).
58. Homes, C. C., Dordevic, S. V., Bonn, D. A., Liang, R. & Hardy, W. N. Sum rules and energy scales in the high-temperature superconductor $YBa_2Cu_3O_{6+x}$. *Phys. Rev. B* **69,** 024514 (2004).
59. Homes, C. C., Dordevic, S. V., Valla, T. & Strongin, M. Scaling of the superfluid density in high-temperature superconductors. *Phys. Rev. B* **72,** 134517 (2005).
60. Tallon, J. L., Cooper, J. R., Naqib, S. H. & Loram, J. W. Scaling relation for the superfluid density of cuprate superconductors: origins and limits. *Phys. Rev. B* **73,** 180504(R) (2006).



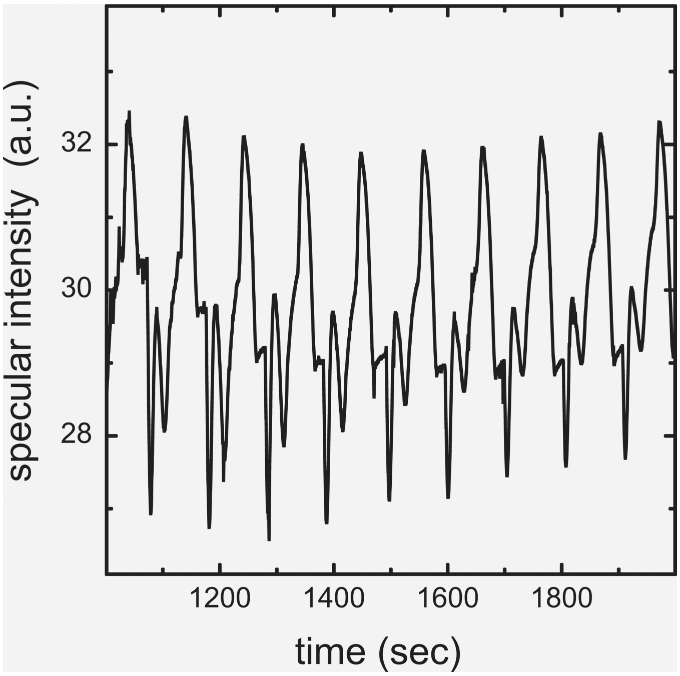
Extended Data Figure 1 | The dependence of T_c on ρ_{s0} in several LSCO films. The films have the same nominal doping near the optimal ($p = 0.16$) in the active (superconducting) layer, but with different thickness $D = nd$, where $d = 0.662$ nm and $n = 1, 2, 4, 10, 40$ and 80. Although individually both T_c and ρ_{s0} show some random variations, in part due to imperfect control of the doping level and the density of the oxygen vacancies, their ratio apparently stays almost constant, to about ±1%. This reinforces the conclusion that T_c is indeed essentially controlled by ρ_{s0}, a purely kinematic quantity.



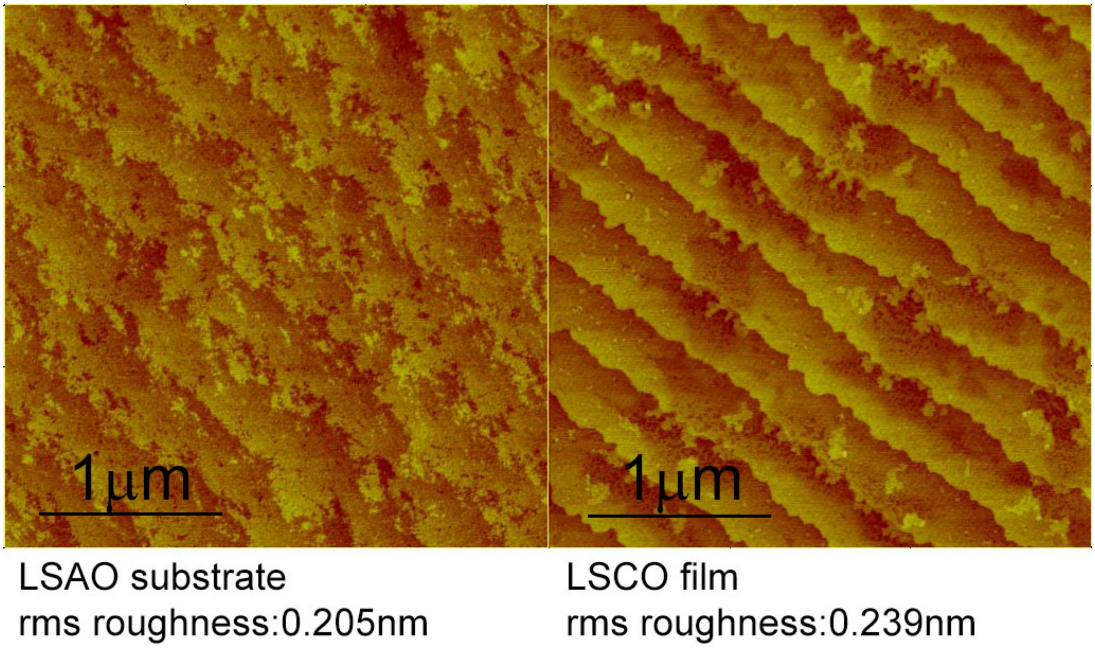
Extended Data Figure 2 | Mutual inductance (raw data) measured on a (275 ± 12)-nm-thick Nb film deposited on standard 10 × 10 × 1 mm^3 LaSrAlO$_4$ substrates. The film was measured on 22 March 2015 (red lines) and 20 January 2016 (black lines). The thermal stabilization is better than ±1 mK and the overall reproducibility is better than ±0.3% on a one-year scale. The inferred value of $\lambda_0 = (41 \pm 5)$ nm agrees with values in the literature[57]; the error here largely comes from the uncertainty in the film thickness. This error is much smaller (down to less than ±1%) in the case of our LSCO films, where we employ atomic-layer deposition, which provides digital control of the film thickness.



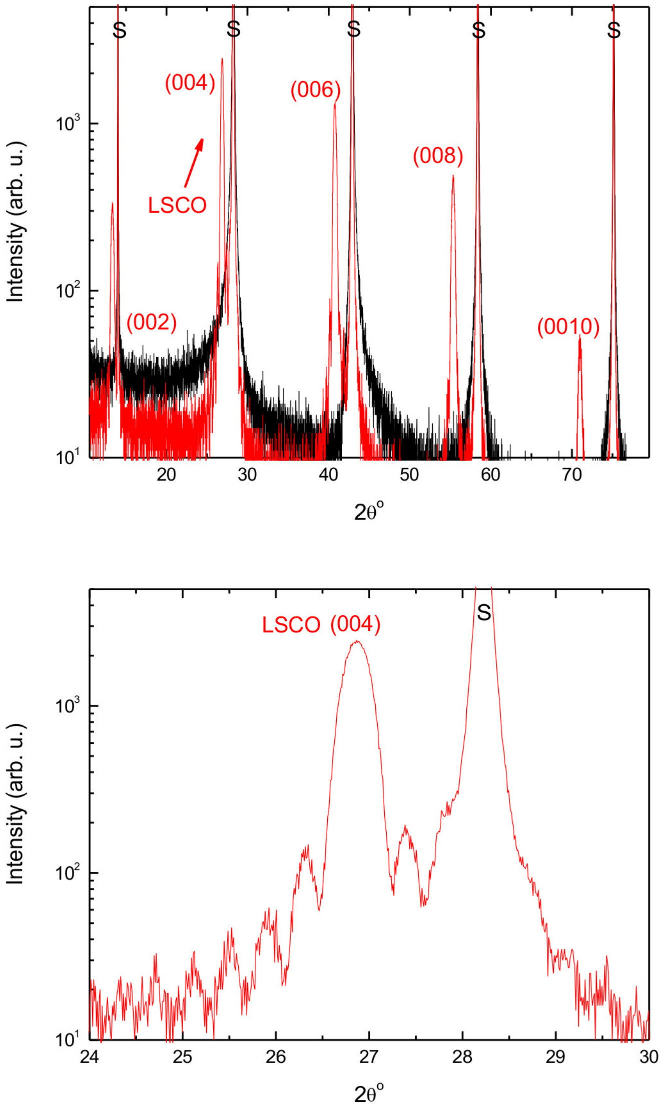
Extended Data Figure 3 | RHEED recorded during growth of LSCO films by ALL-MBE. Top, an optimally doped ($p = 0.16$, $T_c = 40$ K) LSCO film after the end of growth process. Bottom, a strongly overdoped LSCO film ($p = 0.24$, $T_c = 7.5$ K). The stronger main streaks correspond to Bragg-rod reflections at very shallow angles from a terraced surface. The four weaker sidebands in between every pair of main streaks indicate a ubiquitous $5a_0 \times 5a_0$ surface reconstruction (where $a_0 = 0.38$ nm is the in-plane lattice constant). The diagonal streaks are so-called Kikuchi lines that are formed by inelastically scattered electrons; they are observable only from atomically perfect surfaces.



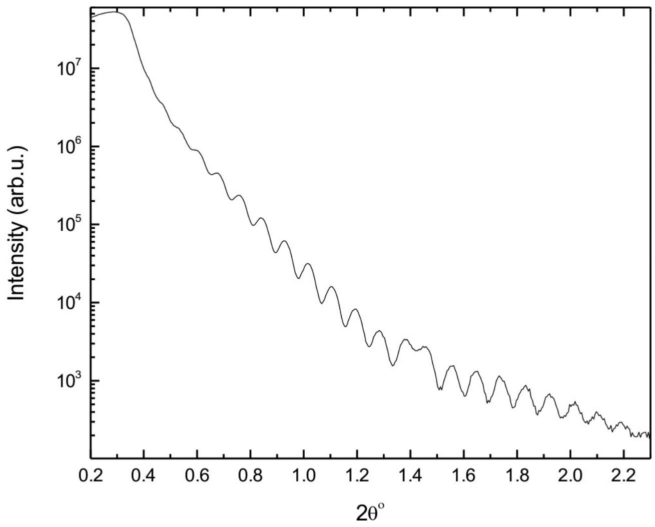
Extended Data Figure 4 | RHEED oscillations recorded during the growth of an LSCO film by ALL-MBE. In the atomic-layer growth mode, the intensity of the specular beam oscillates. When 2D islands form on the surface, the diffuse reflectance increases as the specular reflectance decreases, until about half of the surface is covered. Then the specular reflectance increases again, reaching a new maximum at the full coverage. The fact that the amplitude of the oscillations does not decrease indicates perfect atomic-layer growth. The number of periods provides digital information on the film thickness, expressed in the units of the lattice constant (which we know accurately from XRD).



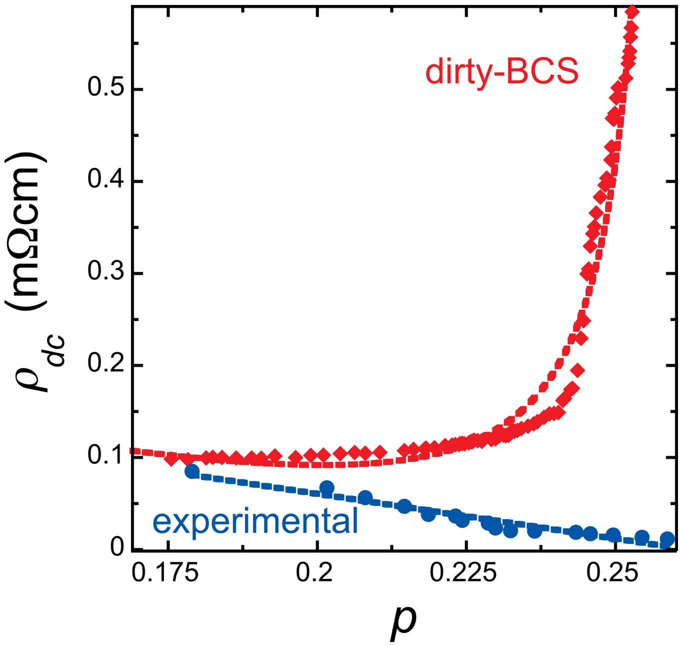
Extended Data Figure 5 | AFM images showing the quality of the film surfaces. Left, an LSAO substrate. The steps, 0.5 UC (0.65 nm) high, occur because the polished surface is unintentionally (but unavoidably) oriented slightly (typically by less than 0.3°) off the desired crystallographic plane perpendicular to the [001] direction. Right, a 225-Å-thick LSCO film grown on the same substrate. The steps in the substrate are projected onto the film and persist all the way to the film surface, indicating atomic-layer growth. The overall root-mean-square (r.m.s.) surface roughness is about 0.24 nm; the terraces between steps are atomically smooth.



Extended Data Figure 6 | A wide-angle 2*θ* XRD pattern of an LSCO film grown on an LSAO substrate by ALL-MBE. The top panel shows a pristine LSAO substrate (black) and an LSCO film grown on the same substrate (red). Only even-order reflections are allowed by the space-group symmetry. The substrate peaks are labelled S. There are no traces of any other phases. The bottom panel is an expanded view near the (004) LSCO reflection. The side-bands between the LSCO Bragg reflections are the so-called Laüe (or finite-thickness) fringes that originate from the interference between X-rays reflected from the film surface and the substrate–film interface.



Extended Data Figure 7 | Low-angle X-ray reflectivity measured from an LSCO film grown on an LSAO substrate by ALL-MBE. The oscillations are so-called Kiesig fringes that originate from interference between X-rays reflected from the film surface and the substrate–film interface. They are analogous to a Fabry–Perot interferogram, and indicate that the two 'mirrors' are smooth and parallel on the scale of the wavelength of light (here, 1.54 Å). By comparing with simulated interferograms, one can estimate the film thickness and roughness; the estimates agree well with the thickness inferred from the digital count of the unit cells by RHEED and the surface roughness as determined by AFM.



Extended Data Figure 8 | Failure of the dirty BCS model to account for experimental data. The red diamonds represent ρ_{dc} calculated from our measured T_{c} and ρ_{s0} values by applying Homes' law, $\rho_{\mathrm{s0}} \propto \sigma_{\mathrm{dc}}T$, which follows from the Ferrell–Glover–Tinkham sum rule for dirty BCS superconductors[23,28,58–60]. As $T_{\mathrm{c}} \propto \sqrt{\rho_{\mathrm{s0}}}$ at high overdoping, the predicted ρ_{dc} value diverges as $1/\sqrt{\rho_{\mathrm{s0}}}$, which should trigger a superconductor-to-insulator transition. The red dashed line is a fit to $f(p) = c_1 + c_2 p + c_3/(0.26 - p)$. Blue circles are the measured ρ_{dc} values (from the data shown in Fig. 3a) showing that the samples in fact get more metallic. The blue dashed line is a fit to $f(p) = c_1 - c_2 p$. The gross discrepancy with the experiment implies that the original premise—the dirty BCS scenario—is incorrect.

doi:10.1038/nature18306

High-efficiency two-dimensional Ruddlesden–Popper perovskite solar cells

Hsinhan Tsai[1,2]*, Wanyi Nie[1]*, Jean-Christophe Blancon[1], Constantinos C. Stoumpos[3,4,5], Reza Asadpour[6], Boris Harutyunyan[4,5], Amanda J. Neukirch[1], Rafael Verduzco[2,7], Jared J. Crochet[1], Sergei Tretiak[1], Laurent Pedesseau[8], Jacky Even[8], Muhammad A. Alam[6], Gautam Gupta[1], Jun Lou[2], Pulickel M. Ajayan[2], Michael J. Bedzyk[4,5], Mercouri G. Kanatzidis[3,4,5] & Aditya D. Mohite[1]

Three-dimensional organic–inorganic perovskites have emerged as one of the most promising thin-film solar cell materials owing to their remarkable photophysical properties[1–5], which have led to power conversion efficiencies exceeding 20 per cent[6,7], with the prospect of further improvements towards the Shockley–Queisser limit for a single-junction solar cell (33.5 per cent)[8]. Besides efficiency, another critical factor for photovoltaics and other optoelectronic applications is environmental stability and photostability under operating conditions[9–15]. In contrast to their three-dimensional counterparts, Ruddlesden–Popper phases—layered two-dimensional perovskite films—have shown promising stability, but poor efficiency at only 4.73 per cent[13,16,17]. This relatively poor efficiency is attributed to the inhibition of out-of-plane charge transport by the organic cations, which act like insulating spacing layers between the conducting inorganic slabs. Here we overcome this issue in layered perovskites by producing thin films of near-single-crystalline quality, in which the crystallographic planes of the inorganic perovskite component have a strongly preferential out-of-plane alignment with respect to the contacts in planar solar cells to facilitate efficient charge transport. We report a photovoltaic efficiency of 12.52 per cent with no hysteresis, and the devices exhibit greatly improved stability in comparison to their three-dimensional counterparts when subjected to light, humidity and heat stress tests. Unencapsulated two-dimensional perovskite devices retain over 60 per cent of their efficiency for over 2,250 hours under constant, standard (AM1.5G) illumination, and exhibit greater tolerance to 65 per cent relative humidity than do three-dimensional equivalents. When the devices are encapsulated, the layered devices do not show any degradation under constant AM1.5G illumination or humidity. We anticipate that these results will lead to the growth of single-crystalline, solution-processed, layered, hybrid, perovskite thin films, which are essential for high-performance opto-electronic devices with technologically relevant long-term stability.

The crystal structures of the Ruddlesden–Popper layered perovskites used in this study, $(BA)_2(MA)_2Pb_3I_{10}$ ($n=3$) and $(BA)_2(MA)_3Pb_4I_{13}$ ($n=4$), are illustrated in Fig. 1a. Both are members of the $(BA)_2(MA)_{n-1}Pb_nI_{3n+1}$ layered perovskite family, where $\{(MA)_{n-1}Pb_nI_{3n+1}\}^{2-}$ denotes the anionic layers derived from the parent 3D perovskite, methylammonium lead triiodide ($MAPbI_3$). The anionic layers are isolated from one another by means of the organic *n*-butylammonium (BA) spacer cations. Within the series, the thickness of each perovskite layer can be incrementally adjusted by careful control of the stoichiometry. Here we focus on the $n=4$ member (Fig. 1a, lower panel) unless otherwise stated.

The layered perovskite thin films were fabricated using a hot-casting technique (see Methods for details)[18]. Briefly, the spin coating of the perovskite precursors was performed on a preheated substrate (fluorine-doped tin oxide (FTO)/poly(3,4-ethylenedioxythiophene) polystyrene sulfonate (PEDOT:PSS)). The temperature of the substrate was varied from room temperature to 150 °C, as illustrated in the images shown in Fig. 1b. The hot-cast films look uniform and reflective, suggesting that the film is ideal for fabrication of planar devices. Atomic force microscopy (AFM) and scanning electron microscopy (SEM) were used to compare the morphologies between room-temperature-cast and hot-cast films (Extended Data Fig. 1a–d). From the AFM images, the hot-cast films show substantially larger grains (about 400 nm) in comparison to the room-temperature spin-coated films (about 150 nm), resulting in a much more compact and uniform thin film. SEM images further confirm the much lower density of pinholes in the hot-cast films in comparison to RT cast films. The presence of pinholes makes it challenging to fabricate working planar-type devices.

We further investigated the crystallinity of the perovskite films using the grazing incidence X-ray diffraction (GIXRD) technique (Fig. 1c). Three dominant planes were observed at diffraction angles (2θ, Cu K_α) of 14.20°, 28.48° and 43.28°, representing the $(BA)_2(MA)_3Pb_4I_{13}$ crystallographic planes $(\bar{1}1\bar{1})$, (202) and (313), respectively, in both of the films[16]. The full-width at half-maximum (FWHM) as a function of hot-casting temperature for the (202) plane is plotted in Fig. 1d; the FWHM of the (202) plane is greatly reduced from 0.63° to 0.29° when the casting temperature increase from room temperature to 110 °C, and remains constant at 0.27° for higher temperatures. The diffraction background and FWHM for each plane are also reduced for the hot-cast film (Fig. 1c and Fig. 1d inset). These results suggest that the crystallinity of the hot-cast films is superior to that of the room-temperature-cast films and becomes optimal at 110 °C. This stark difference in the observed GIXRD patterns motivated in-depth crystallography measurements using grazing incidence wide-angle X-ray scattering (GIWAXS) imaging, and is discussed with reference to Fig. 2.

The optical absorbance and photoluminescence spectra of the perovskite thin films are illustrated in Fig. 1e (see also Extended Data Fig. 2). The photoluminescence and absorption yield band-edge energies of 1.655 ± 0.002 eV and 1.66 ± 0.01 eV, respectively, in good agreement with previous reports[16]. Density functional theory (DFT)

[1]Los Alamos National Laboratory, Los Alamos, New Mexico 87545, USA. [2]Department of Materials Science and Nanoengineering, Rice University, Houston, Texas 77005, USA. [3]Department of Chemistry, Northwestern University, Evanston, Illinois 60208, USA. [4]Department of Materials Science, Northwestern University, Evanston, Illinois 60208, USA. [5]Engineering and Argonne-Northwestern Solar Energy Research (ANSER) Center, Northwestern University, Evanston, Illinois 60208, USA. [6]School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana 47907, USA. [7]Department of Chemical and Biomolecular Engineering, Rice University, Houston, Texas 77005, USA. [8]Fonctions Optiques pour les Technologies de l'Information, FOTON UMR 6082, CNRS, INSA de Rennes, 35708 Rennes, France.
*These authors contributed equally to this work.

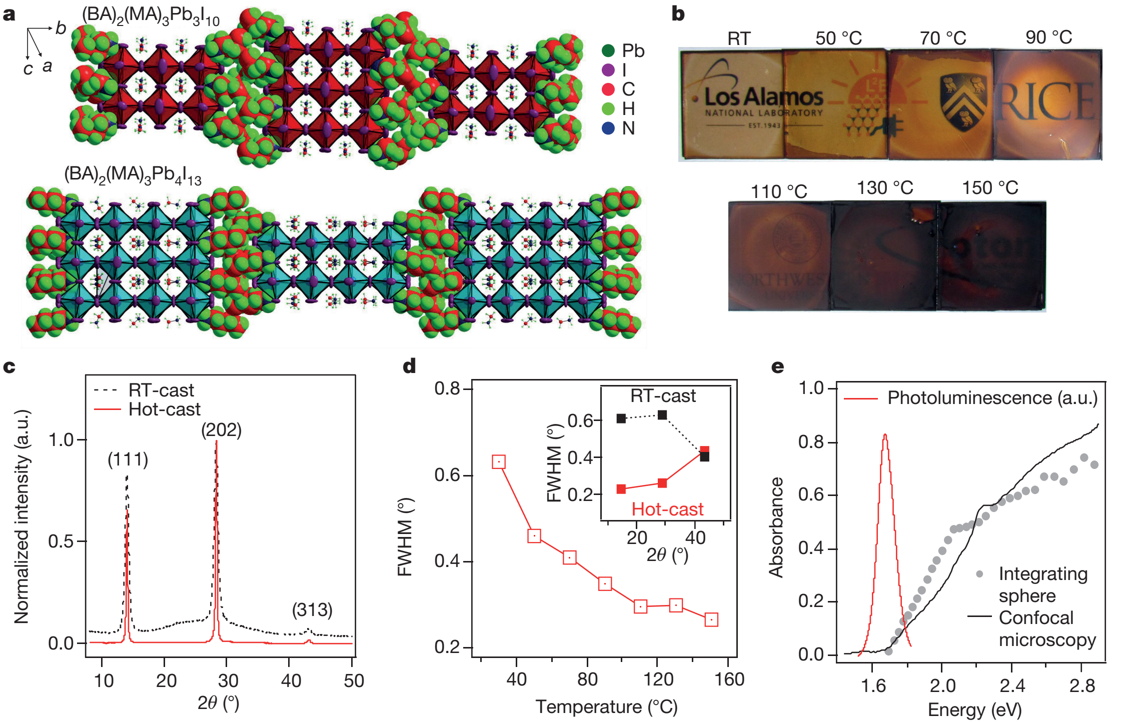


Figure 1 | Crystal structure and thin-film characterization of layered perovskites. a, The crystal structure of the Ruddlesden–Popper $(BA)_2(MA)_2Pb_3I_{10}$ and $(BA)_2(MA)_3Pb_4I_{13}$ layered perovskites, depicted as *n* polyhedral blocks, where *n* refers to the number of layers; the BA spacer layers are depicted as space-fill models to illustrate the termination of the perovskite layers. **b**, Photos of $(BA)_2(MA)_3Pb_4I_{13}$ thin films cast from room temperature (RT) to 150 °C. The film colour gets darker with increasing temperature. **c**, Comparison of GIXRD spectra for room-temperature-cast (black dashed line) and hot-cast (red line) $(BA)_2(MA)_3Pb_4I_{13}$ films, respectively. The (111), (202) and (313) labels correspond to preferred diffraction planes. **d**, The full-width at half-maximum (FWHM) of GIXRD peak (202) as a function of temperature from room temperature to 150 °C. The inset shows the FWHM for each plane indicated in **c**. **e**, Absorbance of a 370-nm thin film ($n = 4$) measured in an integrating sphere (grey circles) and with confocal microscopy (black line), along with the photoluminescence spectra for excitation at 1.96 eV (red line). a.u., arbitrary units.

computations predict that $(BA)_2(MA)_{n-1}Pb_nI_{3n+1}$ compounds have a direct bandgap, with gap energies essentially related to the number of inorganic layers (see Methods, Extended Data Fig. 3 and Supplementary Discussion). The bandgap energy (E_g) can indeed be tuned experimentally from $E_g = 1.52$ eV for $n \to \infty$, similar to 3D $MAPbI_3$, to $E_g = 2.43$ eV for a single atomic layer ($n = 1$), in good agreement with experimental results[12,16,19]. Furthermore, we predict that for $(BA)_2(MA)_2Pb_3I_{10}$ ($n = 3$) and $(BA)_2(MA)_3Pb_4I_{13}$ ($n = 4$) compounds, the exciton binding energy is closer to that of $MAPbI_3$ ($n \to \infty$), for which the excitons are expected to be almost ionized at room temperature and charge-carrier transport is expected to be dominated by free carriers (Supplementary Discussion). These theoretical predictions are in good agreement with the experimentally measured optical absorption spectra, which do not exhibit excitonic signatures (Fig. 1e). Moreover, the apparent lack of Urbach tails in the optical absorption, the very small Stokes shift, and the strong absorption and photoluminescence are indicative that the $(BA)_2(MA)_3Pb_4I_{13}$ perovskite behaves like a direct-bandgap intrinsic semiconductor with excellent crystallinity, very few carrier traps and disorder-induced density of states in the bandgap[20].

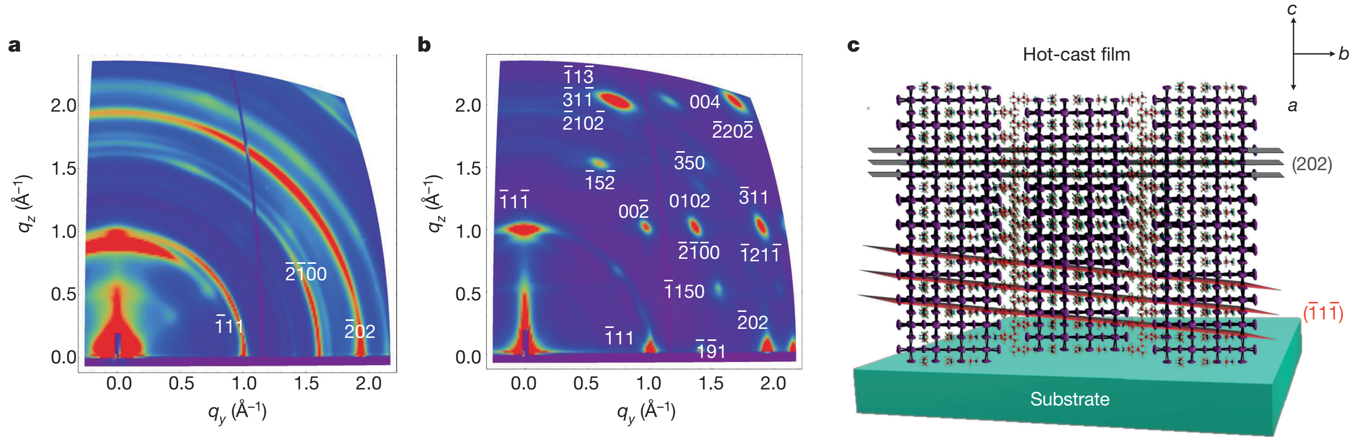


Figure 2 | GIWAXS images and structure orientation. a, **b**, GIWAXS maps for polycrystalline room-temperature-cast (**a**) and hot-cast (**b**) near-single-crystalline $(BA)_2(MA)_3Pb_4I_{13}$ perovskite films with Miller indices of the most prominent peaks shown in white. Colour scale is proportional to X-ray scattering intensity. **c**, Schematic representation of the (101) orientation, along with the $(\bar{1}1\bar{1})$ and (202) planes of a 2D perovskite crystal, consistent with the GIWAXS data.

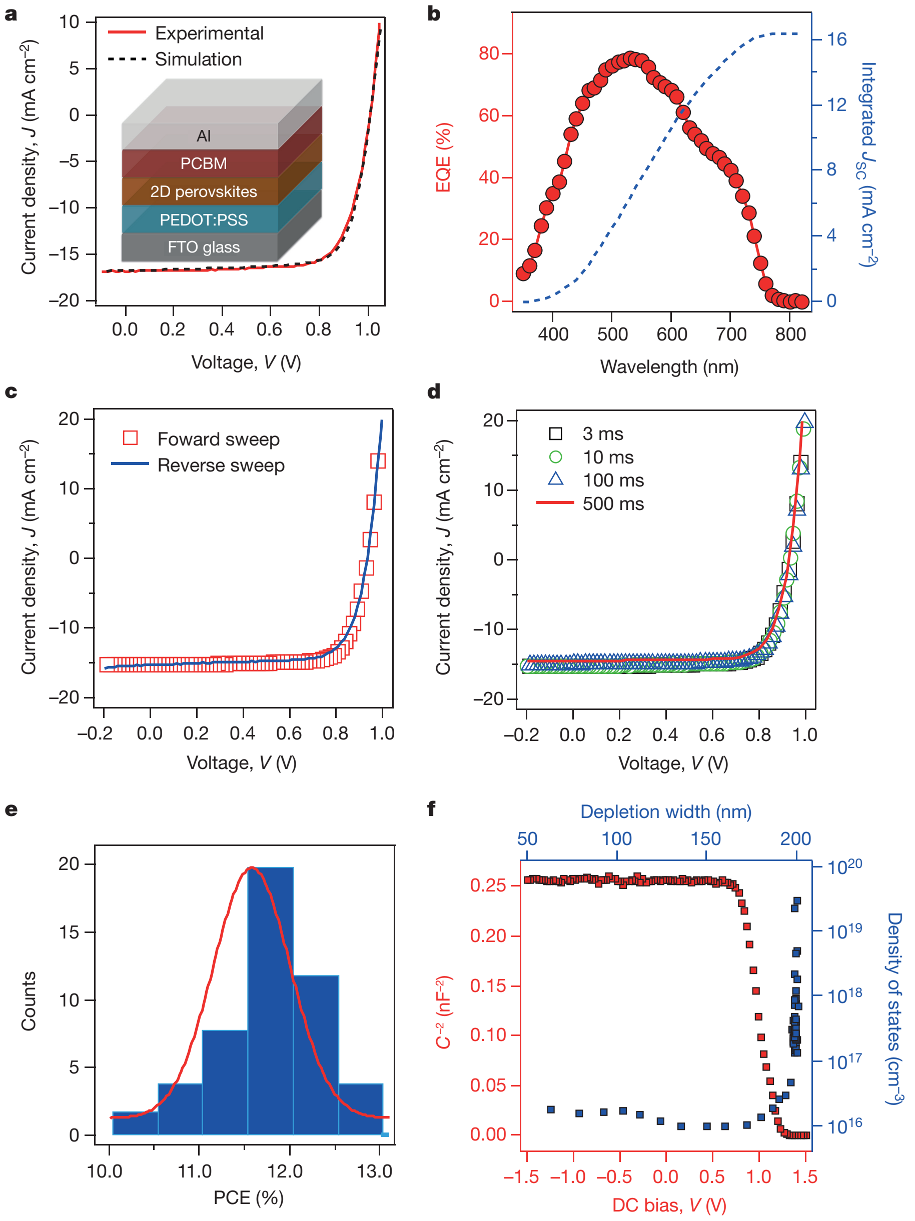


Figure 3 | Solar cell architecture and characterization. a, Experimental (red line) and simulated (black dashed line) current-density–voltage (*J*–*V*) curves under an AM.1.5G solar simulator for planar devices using 2D $(BA)_2(MA)_3Pb_4I_{13}$ perovskites as the absorbing layer at optimized thickness (230 nm). The inset shows the device architecture. Al, aluminium; PCBM, [6,6]-phenyl-C61-butyric acid methyl ester; PEDOT:PSS, poly(3,4-ethylenedioxythiophene) polystyrene sulfonate; FTO, fluorine-doped tin oxide. **b**, External quantum efficiency (EQE; red circles and line) and integrated short-circuit current density (J_{SC}; blue dashed line) as a function of wavelength. **c**, **d**, *J*–*V* curves for hysteresis tests under AM1.5G illumination measured with the voltage scanned in opposite directions (**c**) and with varying voltage delay times (**d**). **e**, Histogram of $(BA)_2(MA)_3Pb_4I_{13}$ device power conversion efficiency (PCE) over 50 measured devices, fitted with a Gaussian distribution (red line). **f**, Capacitance–d.c. bias (*C*–*V*) curves (red squares) for a typical device detected by a small-amplitude a.c. field (peak-to-peak voltage $V_{PP} = 20$ mV) at an a.c. frequency of 100 kHz, and the corresponding charge density profile (blue squares) extracted from the *C*–*V* curve.

To probe the perovskite orientation with respect to the substrate in the thin films we performed a GIWAXS analysis using synchrotron radiation (Fig. 2a, b). The resulting scattering patterns reveal two major characteristics for the films. The room-temperature-cast films (Fig. 2a) exhibit diffraction rings with stronger intensities along certain extended arc segments, which indicates considerable randomness in the 3D orientation of the crystal domains (grains) within the polycrystalline film. By contrast, the hot-cast films (Fig. 2b) exhibit sharp, discrete Bragg spots along the same rings, indicating a textured polycrystalline film in which the crystal domains are oriented with their (101) planes parallel to the substrate surface (as shown in Fig. 2c) and with a 2D in-plane orientational randomness. This crystallographic determination came from indexing the observed Bragg peaks in Fig. 2b using a simulated diffraction pattern from the orthorhombic structure described above. It is apparent from the synchrotron diffraction data that the major perovskite growth direction lies along the (101) plane that is parallel to the q_z direction, as illustrated in Fig. 2c and confirmed by the presence of the $(\bar{1}1\bar{1})$ and (202) spots as the most prominent reflections with q_z.

The growth of highly uniform, layered perovskite thin films with excellent crystallinity and optical properties motivates their use in planar photovoltaic devices. Solar cells fabricated with these thin films yield high power-conversion efficiency with robust and reproducible performance (Fig. 3). The experimental and simulated current density versus voltage characteristics measured under standard AM1.5G (air mass 1.5 global 1-Sun) illumination for the device with $(BA)_2(MA)_3Pb_4I_{13}$ deposited using the hot-casting method[18] are illustrated in Fig. 3a (inset shows the planar device structure). We record a peak power conversion efficiency (PCE) of 12.51% with an open circuit voltage $V_{OC} = 1.01$ V, short-circuit current density $J_{SC} = 16.76$ mA cm^{-2} and fill factor of 74.13% using hot-cast films of 2D $(BA)_2(MA)_3Pb_4I_{13}$ perovskite (see Extended Data Fig. 4 for $(BA)_2(MA)_2Pb_3I_{10}$ device performance). We emphasize that in comparison to 3D perovskites ($V_{OC} \approx 0.7$–0.9 V)[15,18,21,22], the layered 2D perovskites provide an opportunity to achieve high V_{OC}, in a simple planar architecture (see Supplementary Discussion and Extended Data Fig. 1e). For comparison, we fabricated the device using the same materials, but using a conventional spin-coating method, and obtained average efficiencies of approximately 3%–4% (Extended Data Fig. 1f), similar to a previous study[16]. The large increase in efficiency is attributed to the increase in J_{SC} and fill factor. We assign this increase to the enhanced charge transport and mobility (Extended Data Fig. 5) facilitated by the near-perfect vertical orientation of the $\{(MA)_{n-1}Pb_nI_{3n+1}\}^{2-}$ slabs as demonstrated in

Fig. 2c, relative to the FTO substrate. The excellent crystallinity and thin-film uniformity that are realized by the hot-casting technique lead to continuous charge-transport channels that enable the highly mobile photo-generated carriers to travel through the two-dimensional $\{(MA)_{n-1}Pb_nI_{3n+1}\}^{2-}$ slabs across device electrodes, without being blocked by the insulating spacer layers. The measured external quantum efficiency (EQE) and the integrated J_{SC} for $(BA)_2(MA)_3Pb_4I_{13}$ are illustrated in Fig. 3b; the EQE spectrum is in good agreement with the optical absorption profile (Fig. 1e and Extended Data Fig. 2). The integrated J_{SC} from the EQE spectrum is calculated to be 16.31 mA cm^{-2} for $(BA)_2(MA)_3Pb_4I_{13}$, which is within 3% of the J_{SC} measured under AM1.5G standard irradiation. We also performed a film-thickness optimization for solar cell performance using $(BA)_2(MA)_3Pb_4I_{13}$ in the same device geometry by varying the molar concentration of the solution (0.9 M to 0.115 M based on total Pb^{2+}) and found that J_{SC} was optimal for a film thickness of approximately 230 nm. Although thicker films enhance light harvesting, the charge transport limits the overall efficiency (see Extended Data Fig. 6). We also tested these devices for the detrimental hysteresis effect that has been reported in conventional 3D hybrid perovskite[23–25]. Figure 3c, d illustrates the current density as a function of applied voltage, for different sweep directions and voltage delay times, respectively. In contrast to previous reports[13,16], no hysteresis was observed in either case. In fact, these devices demonstrate excellent reproducibility with an average efficiency of 11.60% ± 0.92% over 50 devices, as illustrated through the statistical distribution presented in Fig. 3e.

To rationalize the remarkable reproducibility and the lack of hysteresis of these devices, we performed capacitance–voltage (C–V) measurements and extracted the charge density profile (Fig. 3f) using the Mott–Schottky equation in the reverse bias regime[26–28] (see Supplementary Discussion). From the C–V curve (Fig. 3f, bottom and left axes), it is clear that the device is fully depleted, and the trap density is quite small to substantially affect the J–V characteristics through trapped charges. The device behaves like a p–i–n junction, where layered perovskite acts as an intrinsic semiconductor. The total charge density in the depletion region was calculated to be 10^{16} cm^{-3} (Fig. 3f, top and right axes)[28,29]. Moreover, the calculated edge of the depletion region, where the charge density becomes flat (about 200 nm), is comparable to the film thickness, which indicates that the built-in field is strong enough to extract the charges effectively, leading to a very efficient device. With a fully depleted junction, the C–V curve for the device is not affected by the direction of the voltage scan (see Extended Data Fig. 7a), owing to the lack of trap states, consistent with the observation of hysteresis-free J–V curves.

We performed a self-consistent optoelectronic simulation (involving the solution of Maxwell, Poisson and drift-diffusion equations) to examine the hypothesis of improved material quality when using the hot-cast method. Material parameters are summarized in Supplementary Discussion, with the defect density and mobility being used as the fitting parameters; however, the fitted defect density is in the range of the measured charge density (see Fig. 3f). Extended Data Fig. 8 shows that the simulation results are in good agreement with experimental data, and that they reproduce the general features of the J–V characteristic of both the hot-cast (12.52% efficiency) and room-temperature-cast (4.44% efficiency) solar cells. The increase in efficiency can be attributed to mobility improvement, reduced defect density and improved series resistance.

We performed aggressive long-term stability measurements on the layered 2D and 3D perovskite devices and thin films (Fig. 4). First, we compared the stability of the devices under constant light illumination (AM1.5G), without any encapsulation layer (see Fig. 4a, b and Methods for details). Figure 4a shows the normalized power conversion efficiency as a function of testing time under constant AM1.5G illumination. The PCE of the 3D perovskite device degrades to 40% of its original value within the first 24 h, followed by slow degradation to <10% of its original value over the next 2,250 h (about 94 days). This result is consistent with reports of same device structure and testing conditions[11,15]. In contrast, the 2D device retained 80% of its original PCE after 200 h and slowly degraded to about 70% after another 2,050 h.

Second, we performed a humidity test on the devices by placing the unencapsulated solar cells in a humidity chamber (with a relative humidity of 65%). Figure 4b shows the normalized PCE as a function of testing time, under 65% relative humidity, for the 2D and 3D perovskite devices. The 3D perovskite devices undergo a marked degradation within the first 10 h of exposure. This degradation is expected, owing to the hygroscopic nature of the MA cation and the PEDOT layer. The unencapsulated 2D perovskite devices also show degradation, but at a much slower rate in comparison to the 3D perovskite devices. We speculate that the slower degradation could be due to the long and bulkier hydrophobic organic side group in the 2D perovskite structure that, to a large degree, can prevent (or delay) the direct exposure to moisture and thus increase its threshold against degradation.

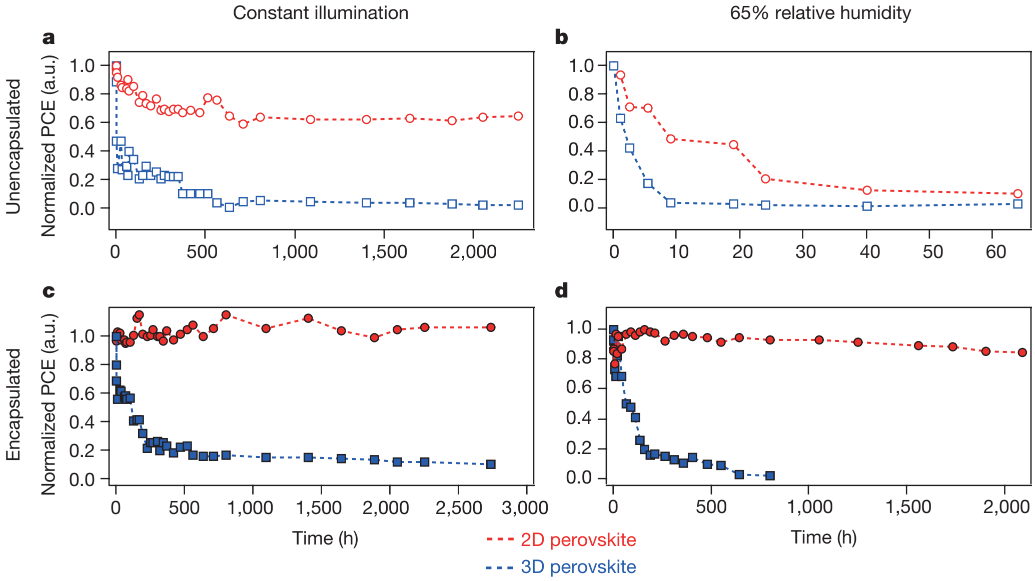


Figure 4 | Stability measurements on planar solar cells. **a**, **c**, Photostability tests under constant AM1.5G illumination for 2D ($(BA)_2(MA)_3Pb_4I_{13}$; red) and 3D ($MAPbI_3$; blue) perovskite devices without (**a**) and with (**c**) encapsulation. **b**, **d**, Humidity stability tests under 65% relative humidity at in a humidity chamber for 2D ($(BA)_2(MA)_3Pb_4I_{13}$; red) and 3D ($MAPbI_3$; blue) perovskite devices without (**b**) and with (**d**) encapsulation. PCE, power conversion efficiency; a.u., arbitrary units.

Finally, we performed the same set of stability tests (under light or humidity) on the 2D and 3D devices with simple glass encapsulation with a UV-curable epoxy resin. Figure 4c, d shows the normalized PCE for encapsulated 2D and 3D perovskite devices tested under constant AM1.5G illumination and 65% humidity for over 2,250 h. Under light stress (Fig. 4c), the PCE degraded to about 50% of its original value within the first 10 h in the 3D device (similar to the unencapsulated 3D device as shown in Fig. 4a), progressively degraded at a very slow rate over the next 800 h and then saturated at approximately 10% of its original value. In contrast, the 2D encapsulated perovskite devices were found to be extremely robust, with no degradation and negligible hysteresis (Extended Data Fig. 7b–d) over 2,250 h of constant light stressing (Fig. 4c). From the light stress tests for the 3D perovskites, we conclude that the observed degradation in the first 10 h in the encapsulated and unencapsulated devices is similar, and so is independent of encapsulation, suggesting that the degradation is a light-activated process.

The humidity stress test for the encapsulated devices (Fig. 4d) reveals that the degradation rate of the PCE for 3D perovskites is much slower than for unencapsulated devices (Fig. 4b), but that the PCE still decreases to <10% of its original value over 350 h. By contrast, the encapsulated 2D devices do not exhibit any degradation in the first 650 h, representing a substantial slowing down of the degradation in humidity.

These stress tests demonstrate that the 2D layered perovskite devices are stable over long-term operation against light soaking and humidity (see heat stress test in Supplementary Discussion and Extended Data Fig. 9), in contrast to 3D perovskite devices. These results reinforce the importance of encapsulation schemes, and motivate the use of design strategies[9,30,31] such as the addition of metal oxide layers onto 3D perovskites to improve charge transport and the encapsulation of layered 2D perovskite devices for long-term stability.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 29 October 2015; accepted 3 May 2016.

Published online 6 July 2016.

1. Yaffe, O. *et al.* Excitons in ultrathin organic-inorganic perovskite crystals. *Phys. Rev. B* **92,** 045414 (2015).
2. Wang, G. *et al.* Wafer-scale growth of large arrays of perovskite microplate crystals for functional electronics and optoelectronics. *Sci. Adv.* **1,** e1500613 (2015).
3. de Quilettes, D. W. *et al.* Impact of microstructure on local carrier lifetime in perovskite solar cells. *Science* **348,** 683–686 (2015).
4. Yin, W.-J., Shi, T. & Yan, Y. Unique properties of halide perovskites as possible origins of the superior solar cell performance. *Adv. Mater.* **26,** 4653–4658 (2014).
5. Stranks, S. D. *et al.* Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber. *Science* **342,** 341–344 (2013).
6. Saliba, M. *et al.* A molecularly engineered hole-transporting material for efficient perovskite solar cells. *Nat. Energy* **1,** 15017 (2016).
7. Yang, W. S. *et al.* High-performance photovoltaic perovskite layers fabricated through intramolecular exchange. *Science* **348,** 1234–1237 (2015).
8. Shockley, W. & Queisser, H. J. Detailed balance limit of efficiency of *p-n* junction solar cells. *J. Appl. Phys.* **32,** 510–519 (1961).
9. You, J. *et al.* Improved air stability of perovskite solar cells via solution-processed metal oxide transport layers. *Nat. Nanotechnol.* **11,** 75–81 (2016).
10. Li, X. *et al.* Improved performance and stability of perovskite solar cells by crystal crosslinking with alkylphosphonic acid ω-ammonium chlorides. *Nat. Chem.* **7,** 703–711 (2015).
11. Kaltenbrunner, M. *et al.* Flexible high power-per-weight perovskite solar cells with chromium oxide–metal contacts for improved stability in air. *Nat. Mater.* **14,** 1032–1039 (2015).
12. Han, Y. *et al.* Degradation observations of encapsulated planar $CH_3NH_3PbI_3$ perovskite solar cells at high temperatures and humidity. *J. Mater. Chem. A Mater. Energy Sustain.* **3,** 8139–8147 (2015).
13. Smith, I. C., Hoke, E. T., Solis-Ibarra, D., McGehee, M. D. & Karunadasa, H. I. A layered hybrid perovskite solar-cell absorber with enhanced moisture stability. *Angew. Chem. Int. Ed.* **53,** 11232–11235 (2014).
14. Noh, J. H., Im, S. H., Heo, J. H., Mandal, T. N. & Seok, S. I. Chemical management for colorful, efficient, and stable inorganic–organic hybrid nanostructured solar cells. *Nano Lett.* **13,** 1764–1769 (2013).
15. Leijtens, T. *et al.* Overcoming ultraviolet light instability of sensitized TiO_2 with meso-superstructured organometal tri-halide perovskite solar cells. *Nat. Commun.* **4,** 2885 (2013).
16. Cao, D. H., Stoumpos, C. C., Farha, O. K., Hupp, J. T. & Kanatzidis, M. G. 2D homologous perovskites as light-absorbing materials for solar cell applications. *J. Am. Chem. Soc.* **137,** 7843–7850 (2015).
17. Kagan, C. R., Mitzi, D. B. & Dimitrakopoulos, C. D. Organic-inorganic hybrid materials as semiconducting channels in thin-film field-effect transistors. *Science* **286,** 945–947 (1999).
18. Nie, W. *et al.* High-efficiency solution-processed perovskite solar cells with millimeter-scale grains. *Science* **347,** 522–525 (2015).
19. Stoumpos, C. C. *et al.* Ruddlesden–Popper hybrid lead iodide perovskite 2D homologous semiconductors. *Chem. Mater.* **28,** 2852–2867 (2016).
20. Sadhanala, A. *et al.* Preparation of single-phase films of $CH_3NH_3Pb(I_{1-x}Br_x)_3$ with sharp optical band edges. *J. Phys. Chem. Lett.* **5,** 2501–2505 (2014).
21. Liang, P.-W. *et al.* Additive enhanced crystallization of solution-processed perovskite for highly efficient planar-heterojunction solar cells. *Adv. Mater.* **26,** 3748–3754 (2014).
22. Jeon, Y.-J. *et al.* Planar heterojunction perovskite solar cells with superior reproducibility. *Sci. Rep.* **4,** 6953 (2014).
23. Tress, W. *et al.* Understanding the rate-dependent *J–V* hysteresis, slow time component, and aging in $CH_3NH_3PbI_3$ perovskite solar cells: the role of a compensated electric field. *Energy Environ. Sci.* **8,** 995–1004 (2015).
24. Unger, E. L. *et al.* Hysteresis and transient behavior in current–voltage measurements of hybrid-perovskite absorber solar cells. *Energy Environ. Sci.* **7,** 3690–3698 (2014).
25. Snaith, H. J. *et al.* Anomalous hysteresis in perovskite solar cells. *J. Phys. Chem. Lett.* **5,** 1511–1515 (2014).
26. Wang, W. *et al.* Device characteristics of CZTSSe thin-film solar cells with 12.6% efficiency. *Adv. Energy Mater.* **4,** 1301465 (2014).
27. Decock, K. *et al.* Defect distributions in thin film solar cells deduced from admittance measurements under different bias voltages. *J. Appl. Phys.* **110,** 063722 (2011).
28. Hegedus, S. S. & Shafarman, W. N. Thin-film solar cells: device measurements and analysis. *Prog. Photovolt. Res. Appl.* **12,** 155–176 (2004).
29. Eisenbarth, T., Unold, T., Caballero, R., Kaufmann, C. A. & Schock, H.-W. Interpretation of admittance, capacitance-voltage, and current-voltage signatures in $Cu(In,Ga)Se_2$ thin film solar cells. *J. Appl. Phys.* **107,** 034509 (2010).
30. Chen, W. *et al.* Efficient and stable large-area perovskite solar cells with inorganic charge extraction layers. *Science* **350,** 944–948 (2015).
31. Mei, A. *et al.* A hole-conductor–free, fully printable mesoscopic perovskite solar cell with high stability. *Science* **345,** 295–298 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work at LANL was LANL LDRD programme (A.D.M., G.G., J.-C.B. and S.T.). Work at Northwestern University was supported as part of the ANSER Center, an Energy Frontier Research Center funded by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Award No. DE-SC0001059. Use of the Advanced Photon Source at Argonne National Laboratory was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under Contract No. DE-AC02-06CH11357. Work at INSA de Rennes was performed using high-performance computational resources from the French national centres (GENCI/CINES/IDRIS grant 2015-c2012096724), Cellule Energie du CNRS (SOLHYBTRANS Project) and the University of Rennes 1 (Action Incitative,Défis Scientifique Emergents 2015). J.E.'s work is also supported by the Fondation d'entreprises banque Populaire de l'Ouest (Grant PEROPHOT 2015). The work at Purdue University was supported by the Bay Area PV Consortium (a Department of Energy project with Prime Award number DE-EE0004946). This work at LANL was done in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT). R.V. acknowledges the support of the NSF DMR-1352099.

Author Contributions A.D.M., H.T., W.N. and M.G.K. conceived the idea. H.T., W.N. and A.D.M. designed the experiments, analysed the data and wrote the paper. H.T. fabricated the devices along with W.N. and performed the measurements. J.-C.B. performed optical spectroscopy measurements and analysed the data under the supervision of J.J.C. C.C.S. synthesized the layered perovskites under the supervision of M.G.K. and co-wrote the paper. R.V. arranged the synchrotron experiments data, and B.H. and M.J.B. analysed and indexed the synchrotron XRD data along with M.G.K. J.E., G.G., L.P. and S.T. performed the molecular dynamics simulations. J.E. analysed the data and provided insight in writing the paper. G.G., J.L. and P.M.A. provided insights into the crystal growth of layered perovskites. R.A. and M.A.A. performed the device simulations. A.J.N. performed DFT calculations on layered perovskites under the supervision of S.T. All authors discussed the results and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.D.M. (amohite@lanl.gov).

Reviewer Information *Nature* thanks H. Snaith and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Materials and instruments. PEDOT:PSS, methylamine hydrochloride (MACl), methylamine solution (MA, 40% in H_2O), hydriodic acid (HI, 57 wt% in H_2O), hypophosphorous acid (H_3PO_2, 50% in H_2O), lead oxide, butylammine (BA, 99%) and *N*,*N*-dimethylformamide (DMF, anhydrous) were purchased from Sigma-Aldrich. All the materials were used as received without further purification. The light source was a simulated AM1.5G irradiance of 100 mW cm^{-2}; calibration was done using a NIST-certified monocrystalline Si solar cell (Newport 532 ISO1599). The scanning electron micrographs were obtained from FEI 400 F with 10 KeV and a spot size of 3.5; AFM images were collected from Bruker Multimode 8.

Materials synthesis. Raw 2D perovskite materials were prepared by combining PbO, MACl and BA in appropriate ratios in a HI/H_3PO_2 solvent mixture as described previously. A detailed experimental procedure is reported in ref. 16.

Materials preparation and device fabrication. Pb_3I_{10} and Pb_4I_{13} were prepared with molar concentrations of 1.8 M, 0.9 M, 0.45 M, 0.225 M and 0.118 M of Pb^{2+} cations in anhydrous DMF. FTO/PEDOT:PSS substrates were prepared following refs 18,32. FTO glasses were cleaned using an ultra-sonication bath in soap water and rinsed progressively with distilled water, acetone and isopropyl alcohol, and finally treated with oxygen plasma for 3 min. The PEDOT:PSS layer was then spin-coated onto the FTO substrates at 5,000 r.p.m. for 45 s as a hole-transporting layer. The coated substrates were then transferred to an argon-filled glovebox for device fabrication. The 2D perovskite solution was prepared by dissolving 0.025 mM 2D perovskite single crystal in DMF. The solution was then heated under continuous stirring at 70 °C for 30 min before device fabrication. For the film hot-casting process, the FTO/PEDOT:PSS substrates were first preheated from 30 °C to 150 °C on a hot plate for 10 min, right before spin-coating. These were immediately (within 5 s) transferred to the hot FTO/PEDOT:PSS substrates on the spin-coated 'chunk' (which is at room temperature), and 80 μl of precursor solution was dropped onto the hot substrate. The spin-coater was immediately started with a spin speed of 5,000 r.p.m. for 20 s without ramp; the colour of the thin film turned from pale yellow to brown in few seconds as the solvent escaped. After the spin-coater stopped, the substrates were quickly removed from it. The [6,6]-phenyl-C61-butyric acid methyl ester (PCBM) solution was prepared by dissolving 20 mg PCBM in 1 ml chlorobenzene. 50 μl of the PCBM solution was then dropped onto the perovskite-coated FTO/PEDOT substrate and spin-coated at 1,000 r.p.m. for 60 s to form a thin, electron-transporting layer. The metal electrodes (Al and Au) were deposited using a thermal evaporator with a shadow mast with a working area of 0.5 cm^2. GIWAXS measurements were carried out on Sector 8-ID-E at the Advanced Photon Source, Argonne National Laboratory.

Derivation of the perovskite orientation from GIWAXS. The unit vector $\hat{\boldsymbol{n}}$ normal to substrate surface was defined in the lattice reference frame. Finding the orientation of the crystallites (which are rotationally random around $\hat{\boldsymbol{n}}$) amounts to finding the Miller indices of the plane perpendicular to $\hat{\boldsymbol{n}}$. Two peaks were chosen whose possible set of indices can be inferred from powder diffraction patterns. For each peak of the pair (peaks at q_y and q_z directions), an equations was set: $\frac{\hat{\boldsymbol{n}}\cdot\boldsymbol{G}}{G} = \cos(\theta) \equiv \frac{q_z}{q}$, where $\boldsymbol{G}$ is the reciprocal lattice vector for the chosen peak and θ is the angle between the total wavevector transfer $\boldsymbol{q}$ and its z-axis component q_z (both deduced from peak positions on the GIWAXS pattern, q through the relation $q = \sqrt{q_x^2 + q_y^2 + q_z^2}$). The obtained system of two equations is solved for the Miller indices of the plane, which has as its normal the $\hat{\boldsymbol{n}}$ unit vector. The peak pair that produced the GIWAXS pattern simulation (performed using the Laue condition $\boldsymbol{G} = \boldsymbol{q}$) best matching the experimental pattern was taken as the optimal solution. The Miller plane parallel to the substrate was deduced to be the (101) plane.

Mobility measurement. We measured the mobility using the charge extraction by linearly increasing voltage (CELIV) technique for the hot-cast and room-temperature-cast devices in the same device geometry. The device is connected to a function generator from the FTO side for negative bias with increasing magnitude, and the current transient is recorded by measuring the voltage drop through a resistor (50 Ω) connected in series with the cathode side of the device.

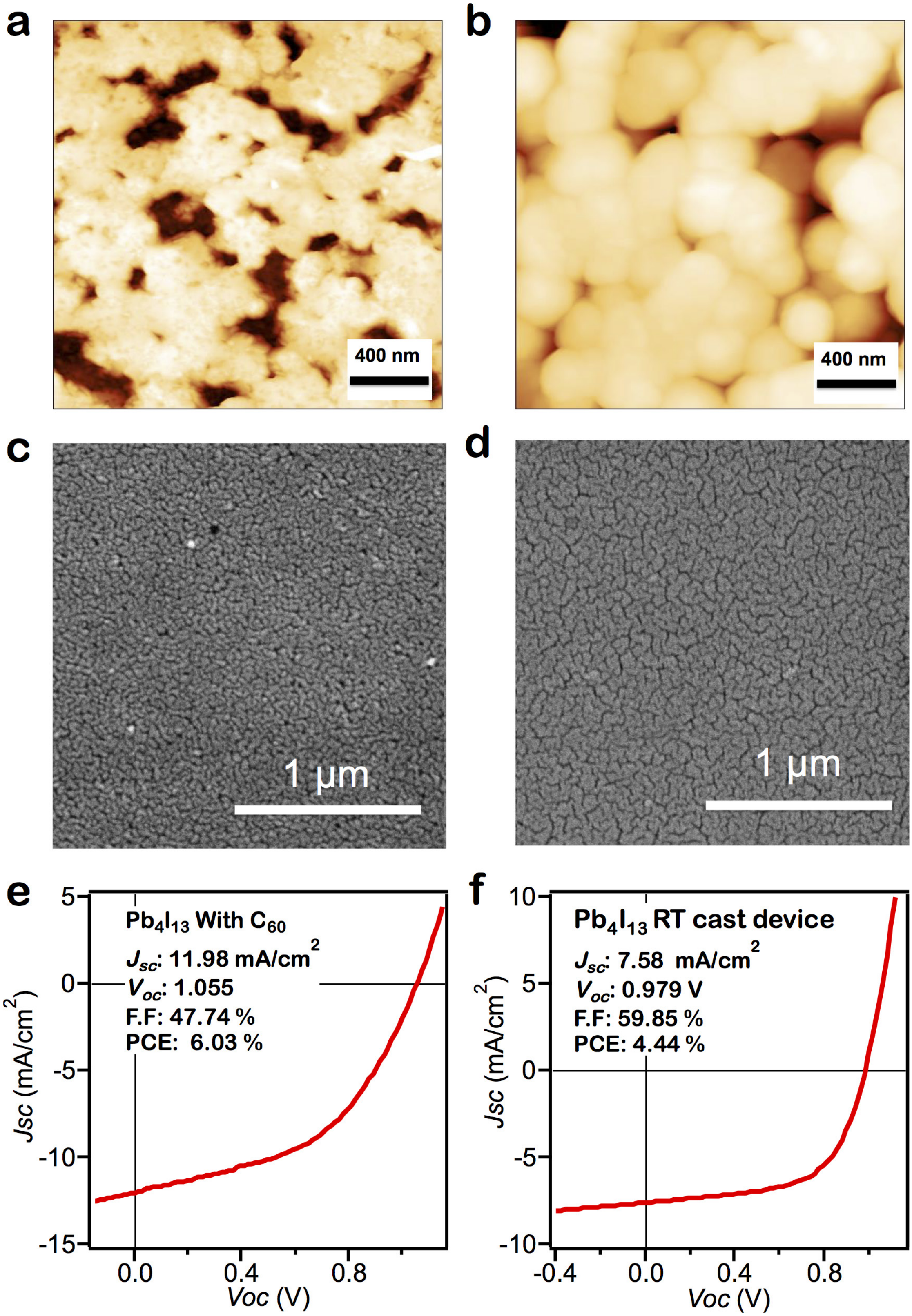
***C*–*V* measurement.** We performed *C*–*V* measurements on the 2D layered perovskite device to characterize the its charge density profile, as shown in Fig. 3f, in which C^{-2} is plotted against d.c. voltage. The corresponding charge density profile was extracted from Fig. 4 using the standard Mott–Schottky equation[33]:

$$C^{-2} = \frac{2(V_{\mathrm{bi}} - V)}{A^2 q \varepsilon_0 \varepsilon N}$$

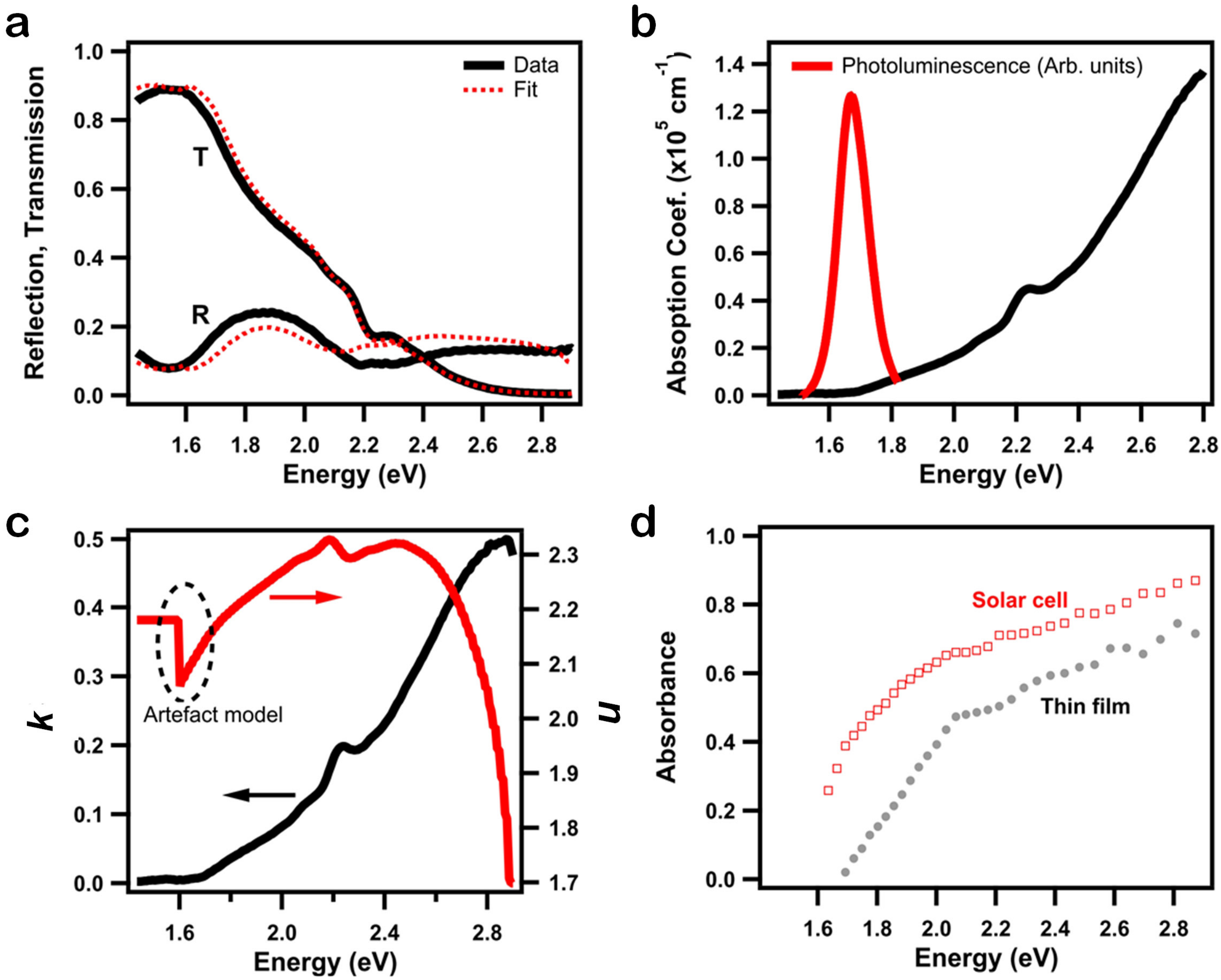
in which V_{bi} is the built-in voltage, A is the device area, q is the elementary charge, ε is the dielectric constant, ε_0 is the vacuum permittivity and N is the charge density. The charge density profile is extracted from the voltage range 0.5 V to −1 V (reverse bias regime), in which there is no d.c. carrier injection and the device behaves as a capacitor. To understand the origin of hysteresis in these devices, we performed *C*–*V* measurements on the 2D devices and measured the charge density profile of the perovskite depletion region in the reverse bias regime from −1.0 V to 0 V (low dark current without charge injection) using the Mott–Schottky equation[33].

Environmental stability test. We first prepared the 2D and 3D perovskites devices in the same device configuration as described in inset of Fig. 3a, with and without encapsulation, by sealing the active area with glass and ultraviolet curable epoxy. For light stress tests, the 2D and 3D perovskite devices were placed under constant AM1.5G illumination, but were taken out from under the light for each *J*–*V* curve scan. For humidity stress tests, the 2D and 3D perovskite devices were placed inside the humidity chamber connected with a desiccant and a humidifier to control the relative humidity (65%); the relative humidity values were calibrated with two digital humidity sensors at two corners of the chamber.

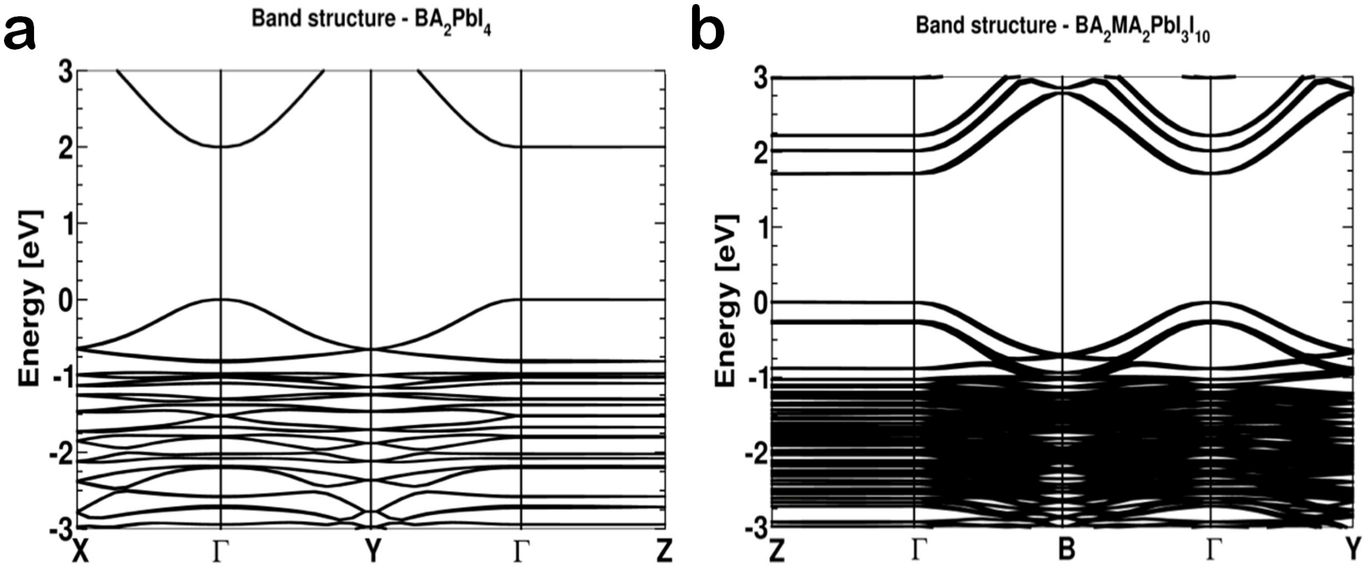
32. Tsai, H. *et al.* Optimizing composition and morphology for large-grain perovskite solar cells via chemical control. *Chem. Mater.* **27,** 5570–5576 (2015).
33. Gelderman, K., Lee, L. & Donne, S. W. Flat-band potential of a semiconductor: using the Mott–Schottky equation. *J. Chem. Educ.* **84,** 685–688 (2007).
34. Schulz, P. *et al.* Interface energetics in organo-metal halide perovskite-based photovoltaic cells. *Energy Environ. Sci.* **7,** 1377–1381 (2014).
35. Brivio, F., Walker, A. B. & Walsh, A. Structural and electronic properties of hybrid perovskites for high-efficiency thin-film photovoltaics from first-principles. *APL Mater.* **1,** 042111 (2013).
36. Trukhanov, V. A., Bruevich, V. V. & Paraschuk, D. Y. Effect of doping on performance of organic solar cells. *Phys. Rev. B* **84,** 205318 (2011).



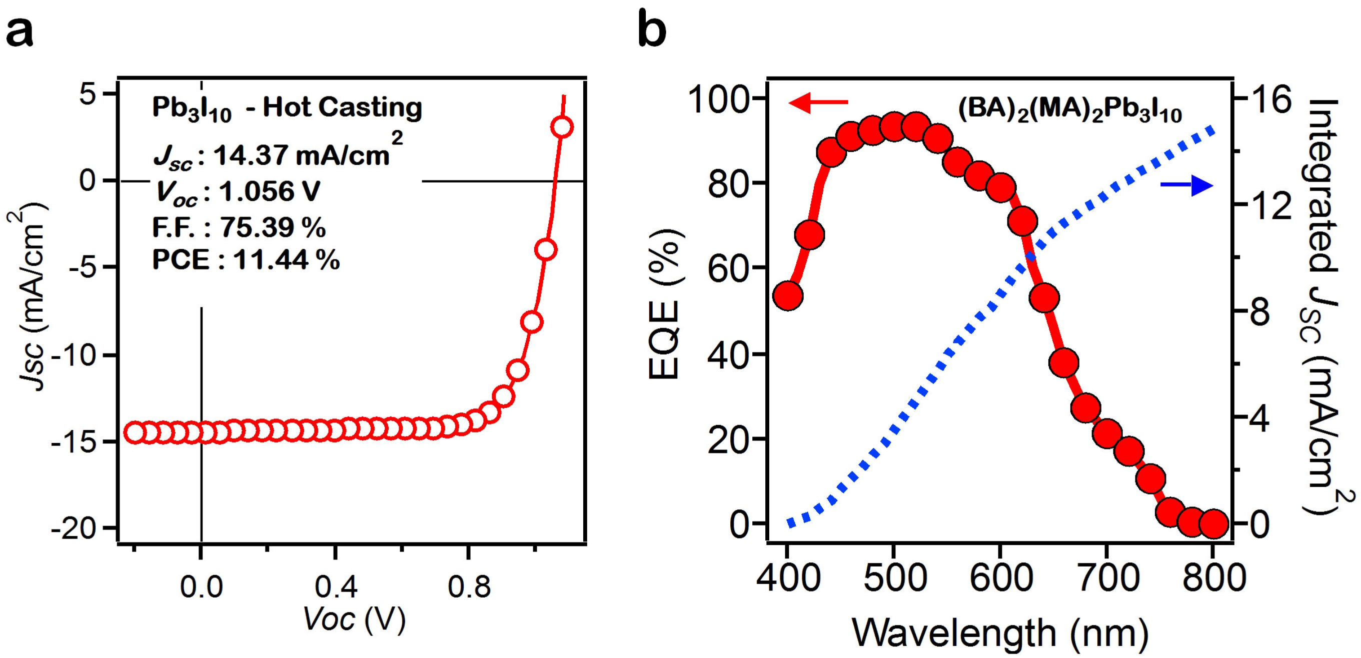
Extended Data Figure 1 | Layered perovskite thin-film morphology and device performance. a, **b**, AFM images of surface morphology for room-temperature-cast (**a**) and hot-cast (**b**) films. Scale bars, 400 nm. **c**, **d**, SEM images of topography for room-temperature-cast (**c**) and hot-cast (**d**) films. Scale bars, 1 µm. **e**, *J*–*V* curve of Pb_4I_{13} with C_{60} as a contact modification candidate shows the enhancement of V_{OC} from 0.9 V to 1.055 V with the same device architecture. **f**, *J*–*V* curve for the $(BA)_2(MA)_3Pb_4I_{13}$ device using the room-temperature (RT) spin-cast method. FF, fill factor.



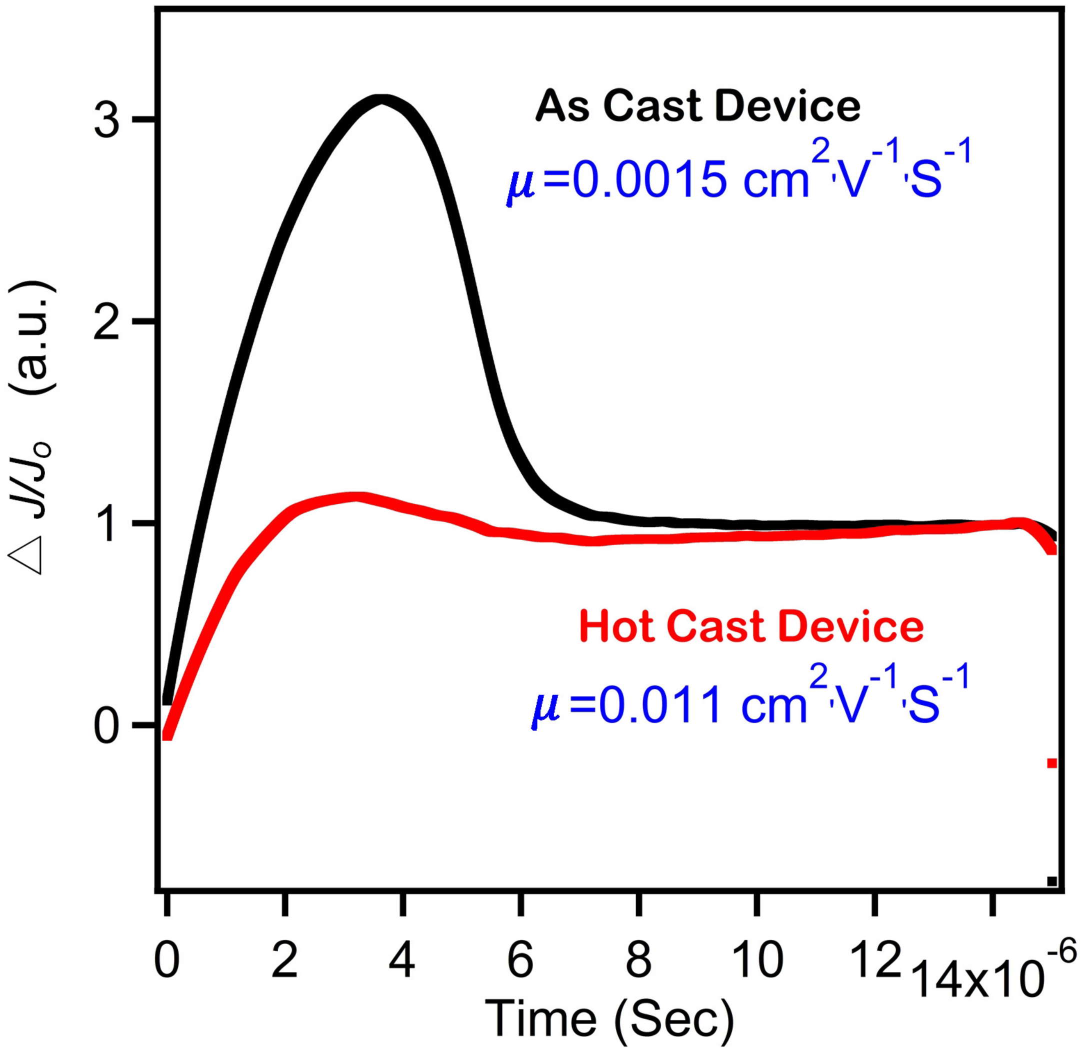
Extended Data Figure 2 | Absorption spectroscopy of layered 2D perovskites. a–**c**, Local optical absorption characteristics of thin films using reflection/transmission experimental methods (see also refs 34,35 for details of the modelling): results of the fitting of the reflection (R) and transmission (T) data (**a**); absolute absorption cross-section (**b**); and real (*n*; red line) and imaginary (*k*; black line) parts of the refractive index (**c**). **d**, Absorbance of thin films (grey circles) compared to that of optimized solar cells (red squares) measured using integrating sphere techniques (see details in ref. 36).



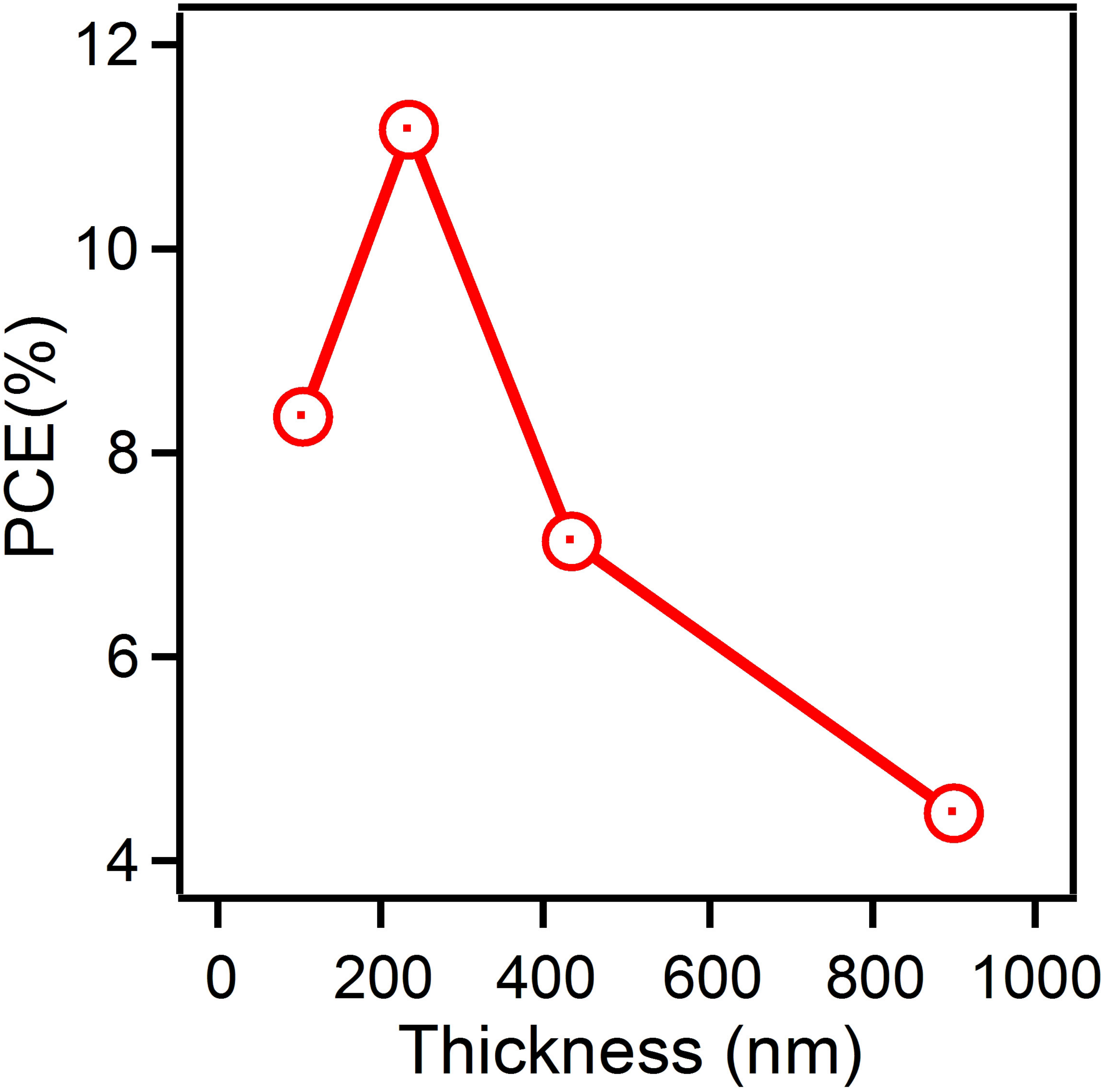
Extended Data Figure 3 | DFT computation. a, **b**, Electronic band structures of $(BA)_2PbI_4$ ($n = 1$; **a**) and $(BA)_2(MA)_2Pb_3I_{10}$ ($n = 3$; **b**) calculated using DFT with a local-density approximation, including the spin–orbit coupling and a bandgap correction computed using the HSE (Heyd–Scuseria–Ernzerhof) functional. The energy levels are referenced to the valence band maximum.



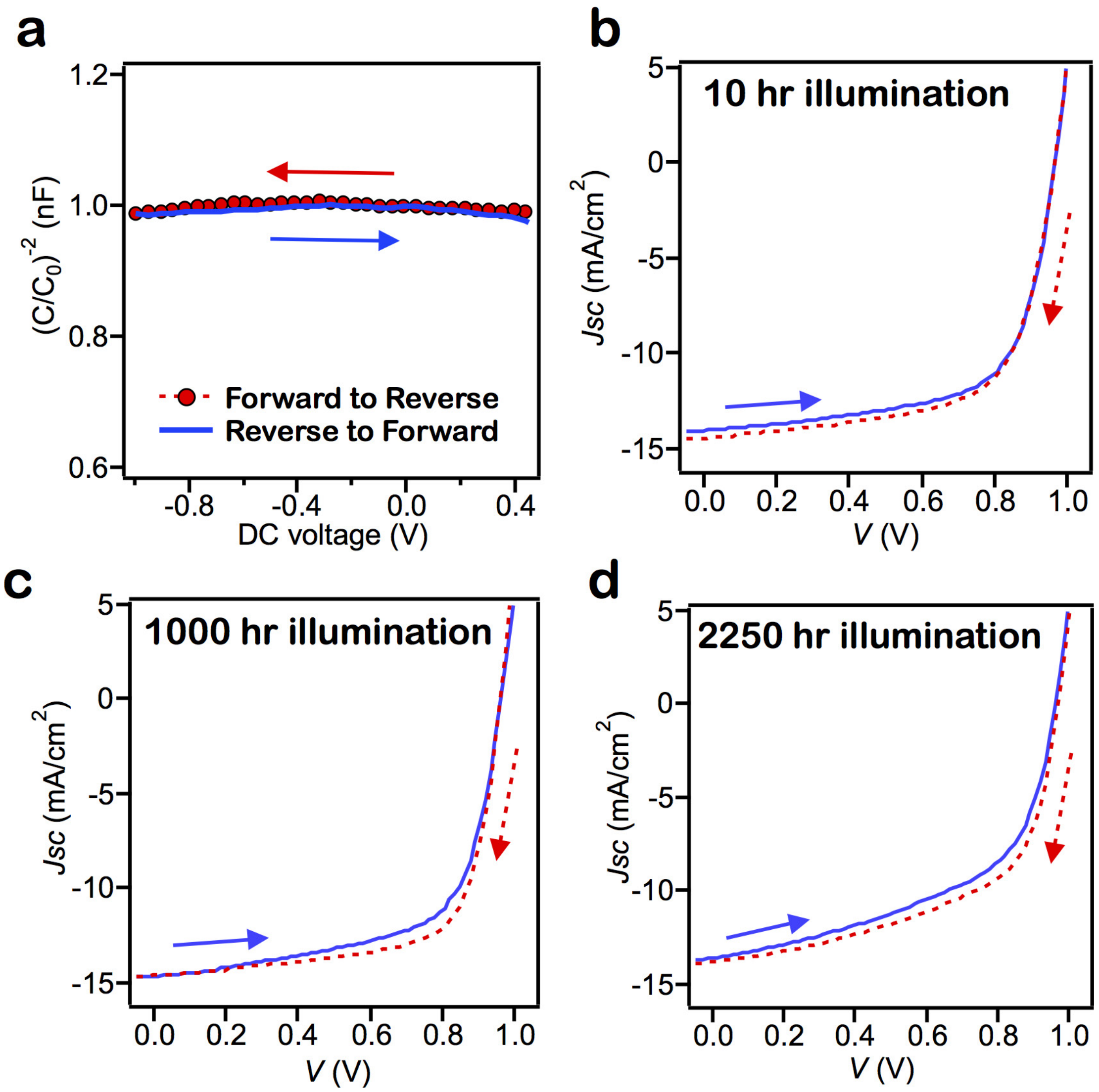
Extended Data Figure 4 | Device performance of $(BA)_2(MA)_2Pb_3I_{10}$. **a**, *J*–*V* curve and device parameters. **b**, EQE (red circles) and integrated J_{SC} from EQE (blue dashed line).



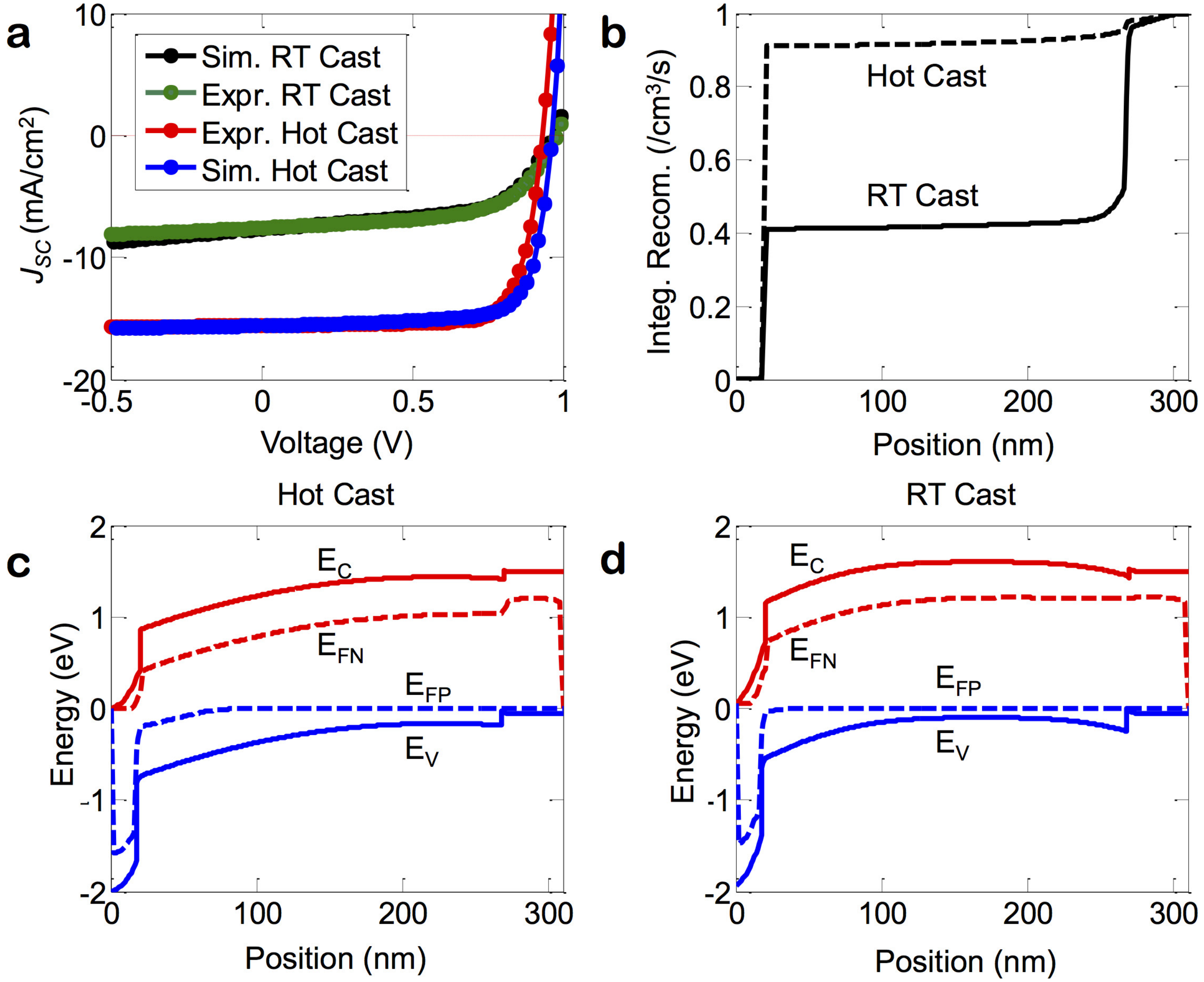
Extended Data Figure 5 | Dark current transient and mobility. The dark current transient ($\Delta J/J_0$), measured using the CELIV technique, for a hot-cast (red) and a room-temperature-cast ('As cast', black) device, and the mobility value (μ) in each case.



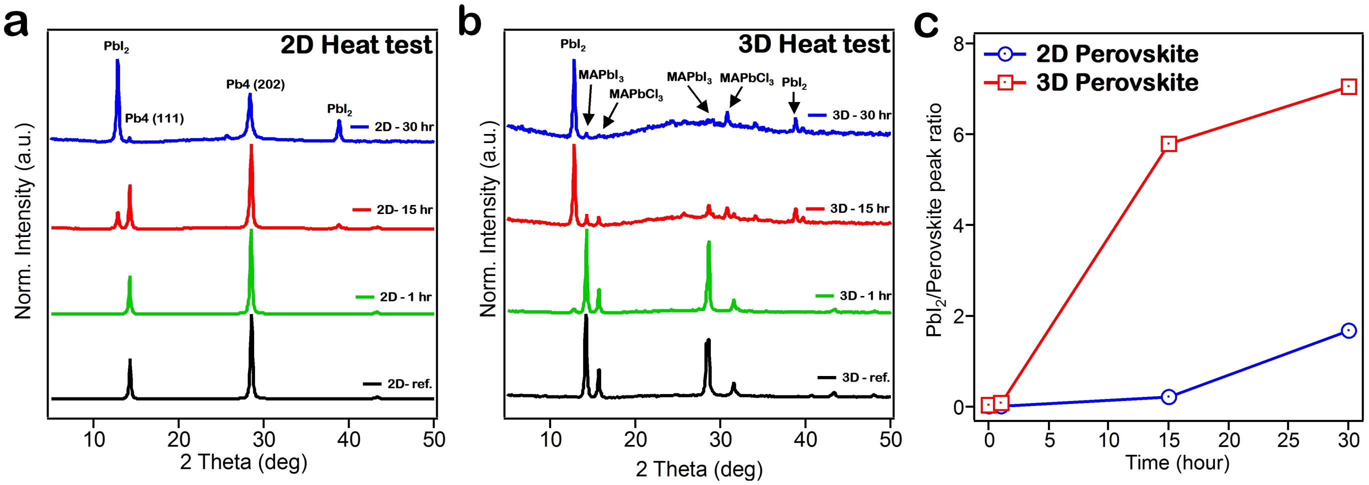
Extended Data Figure 6 | Device PCE as a function of thin-film thickness for the layered Pb_4I_{13} perovskite.



Extended Data Figure 7 | Hysteresis tests for 2D pervoskite devices. a–d, Tests with different bias sweep directions (**a**; $(C/C_0)^{-2}$ as function of DC bias, where C_0 is the capacitance of a geometric capacitor), and after 10 h (**b**), 1,000 h (**c**) and 2,250 h (**d**) of constant illumination. . The red and blue arrows indicate the forward and reverse sweep directions.



Extended Data Figure 8 | Simulation results and comparison of room-temperature-cast and hot-cast methods. a, Experimental ('Expr.') *J*–*V* characteristics of room-temperature-cast ('As cast') and hot-cast methods and corresponding simulation ('Sim.') results. The hot-cast method shows a current density with a larger magnitude and higher fill factor (area below the *J*–*V* curve). **b**, Integrated recombination inside three layers of a solar cell. Peak recombination shifts toward the PCBM/perovskite interface because the barrier for generated carriers is less in the hot-cast case than in the room-temperature-cast case. **c**, **d**, Energy band diagram of hot-cast (**c**) and room-temperature-cast (**d**) methods. Generated carriers face a lower barrier in the hot-cast case, especially close to the PEDOT/perovskite interface. E_C, conduction band; E_V, valence band; E_{FN}, electron quasi-Fermi level; E_{FP}, hole quasi-Fermi level.



Extended Data Figure 9 | Heat stress tests. a, **b**, Spectra of 2D (**a**) and 3D (**b**) perovskite thin films under 80 °C in darkness after the lengths of time indicated (spectra are offset for clarity; 'ref.' refers to freshly made thin film, measured after 0 h of heat stressing). **c**, Ratio of the PbI_2 ($2\theta = 12.7°$) and perovskite ($2\theta = 14.2°$) main peaks in the spectra in **a** and **b** for the two perovskite materials (2D, blue; 3D, red) over 30 h of heating at 80 °C.

doi:10.1038/nature19059

The active site of low-temperature methane hydroxylation in iron-containing zeolites

Benjamin E. R. Snyder[1]*, Pieter Vanelderen[1,2]*, Max L. Bols[2], Simon D. Hallaert[3], Lars H. Böttger[1], Liviu Ungur[3]†, Kristine Pierloot[3], Robert A. Schoonheydt[2], Bert F. Sels[2] & Edward I. Solomon[1,4]

An efficient catalytic process for converting methane into methanol could have far-reaching economic implications. Iron-containing zeolites (microporous aluminosilicate minerals) are noteworthy in this regard, having an outstanding ability to hydroxylate methane rapidly at room temperature to form methanol[1–3]. Reactivity occurs at an extra-lattice active site called α-Fe(ii), which is activated by nitrous oxide to form the reactive intermediate α-O[4,5]; however, despite nearly three decades of research[5], the nature of the active site and the factors determining its exceptional reactivity are unclear. The main difficulty is that the reactive species—α-Fe(ii) and α-O—are challenging to probe spectroscopically: data from bulk techniques such as X-ray absorption spectroscopy and magnetic susceptibility are complicated by contributions from inactive 'spectator' iron. Here we show that a site-selective spectroscopic method regularly used in bioinorganic chemistry can overcome this problem. Magnetic circular dichroism reveals α-Fe(ii) to be a mononuclear, high-spin, square planar Fe(ii) site, while the reactive intermediate, α-O, is a mononuclear, high-spin Fe(iv)=O species, whose exceptional reactivity derives from a constrained coordination geometry enforced by the zeolite lattice. These findings illustrate the value of our approach to exploring active sites in heterogeneous systems. The results also suggest that using matrix constraints to activate metal sites for function—producing what is known in the context of metalloenzymes as an 'entatic' state[6]—might be a useful way to tune the activity of heterogeneous catalysts.

No spectroscopic feature of α-Fe(ii) has thus far been discovered[7]. We have now identified Fe(ii) ligand-field bands in the diffuse reflectance ultraviolet–visible (DR-UV-vis) spectra of three iron-containing zeolites (Fe-zeolites) that are known to contain α-Fe(ii) (see Extended Data Fig. 1)[4,8–10]. Of these, we chose the Fe(ii)-beta (BEA) zeolite for further study because of the higher intensity of its ligand-field bands (see Extended Data Fig. 2 for the structure of the BEA lattice). The DR-UV-vis spectrum of Fe(ii)-BEA (Fig. 1a) is characterized by an intense band at 40,000 cm^{-1}, and three weak ligand-field bands at 15,900 cm^{-1}, 9,000 cm^{-1} and <5,000 cm^{-1} (a shoulder). To determine which of these bands correlates with α-Fe(ii), we activated the sample with nitrous oxide (N_2O; Fig. 1b) and then reacted it with methane (CH_4) at room temperature (Fig. 1c). During activation by N_2O, the 15,900 cm^{-1} band of Fe(ii)-BEA is replaced by a new feature of similar intensity at 16,900 cm^{-1}, along with a weak feature at around 5,000 cm^{-1}. The 16,900 cm^{-1} and 5,000 cm^{-1} bands present after N_2O activation disappear upon reaction with CH_4; we therefore assign these bands to α-O. These are the first absorption features to be conclusively attributed to α-O on the basis of their reactivity to CH_4. The 15,900 cm^{-1} band present before N_2O activation is assigned to α-Fe(ii). The 5,000–13,000 cm^{-1} region of the CH_4-reacted spectrum overlaps with that of Fe(ii)-BEA, indicating that features in this region originate from inactive 'spectator' sites.

The 15,900 cm^{-1} band of α-Fe(ii) is an interesting spectral feature, as it is unusual to observe Fe(ii) ligand-field bands in this high-energy region. Correlation of Fe(ii) sites with hard O/N donors suggests that a band in this region is characteristic of square-planar sites with a spin (*S*) of 2 (refs 11–15). Importantly, the 15,900 cm^{-1} band is also a spectroscopic handle that enables selective study of α-Fe(ii) by variable-temperature variable-field magnetic circular dichroism (VTVH-MCD) spectroscopy.

Low-temperature magnetic circular dichroism (MCD) data from Fe(ii)-BEA stabilized in perfluorocarbon glass are shown in Fig. 2a, along with ligand-field DR-UV-vis data for comparison. On the basis of the DR-UV-vis results, the positive MCD feature at 15,100 cm^{-1}

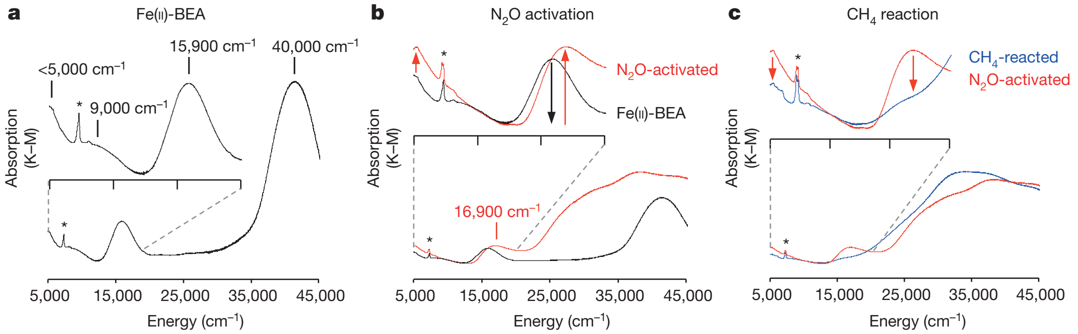


Figure 1 | DR-UV-vis spectra of Fe-BEA in Kubelka–Munk units[31]. **a**, Spectrum for Fe-BEA (Si/Al = 12, 0.3 wt% Fe) after treatment with helium (900 °C, 2 hours) and reduction with hydrogen (700 °C, 1 hour) to produce Fe(ii)-BEA. The overall spectrum is shown at the bottom, with an expanded portion of it above. * = OH overtone. **b**, **c**, The spectra resulting from activation of Fe(ii)-BEA with N_2O at 250 °C for 15 minutes (**b**), followed by reaction with CH_4 at room temperature (**c**). Key spectral changes are indicated with arrows.

[1]Department of Chemistry, Stanford University, Stanford, California 94305, USA. [2]Department of Microbial and Molecular Systems, Centre for Surface Chemistry and Catalysis, KU Leuven – University of Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium. [3]Department of Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium. [4]Photon Science, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, California 94025, USA. †Present address: Division of Theoretical Chemistry, Lund University, PO Box 124, 221 00 Lund, Sweden.
*These authors contributed equally to this work.

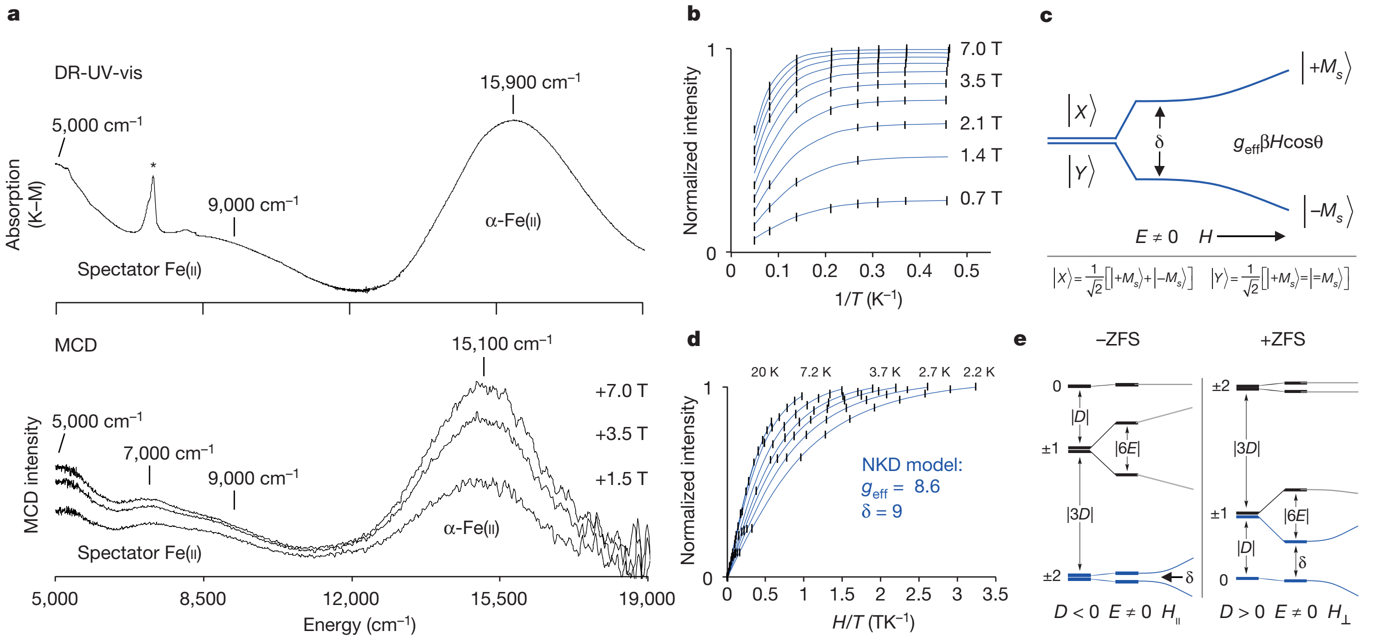


Figure 2 | MCD and VTVH-MCD of α-Fe(II). **a**, Comparison of the 298 K DR-UV-vis data (reproduced from Fig. 1a; * = OH overtone) and variable-field 3 K MCD data from Fe(II)-BEA (Si/Al = 12, 0.3 wt% Fe). **b**, Saturation magnetization isofields for the 15,100 cm^{-1} band (± 1σ error bars are in black; the fit is in blue). T, tesla; T, temperature. **c**, Influence of rhombic zero-field splitting (ZFS) and magnetic field (parameterized by δ and g_{eff}, respectively) on a non-Kramers doublet (NKD). At zero field, the NKD is split in energy by the rhombic ZFS. The $\pm M_s$ wavefunctions also mix equally, to form $|X\rangle$ and $|Y\rangle$. Application of a magnetic field further splits the levels, and changes the wavefunctions to pure $|-M_s\rangle$ and $|+M_s\rangle$ in high fields. **d**, Saturation magnetization isotherms for α-Fe(II) (fit in blue using the NKD model in **c**). **e**, Comparison of +ZFS and –ZFS $S = 2$ spin manifolds, including effects due to axial ($D \neq 0$) and rhombic ($E \neq 0$) ZFS. Levels identified with the non-Kramers doublet model shown in **c** are highlighted in blue.

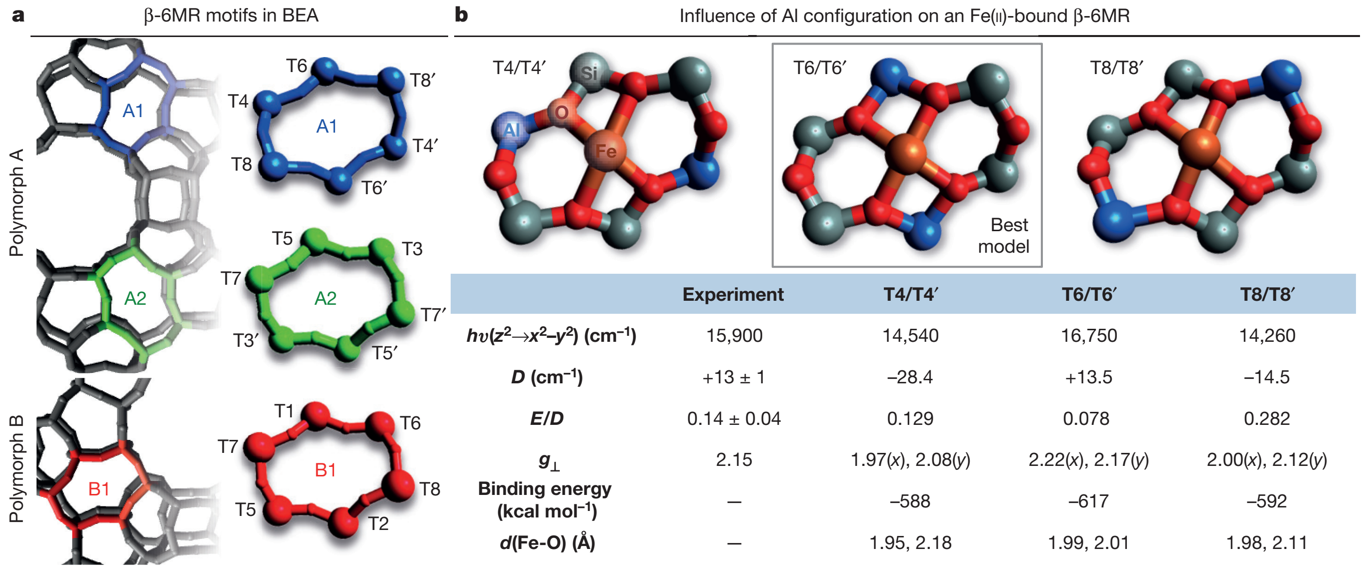
is correlated with α-Fe(II), while the weak positive features between 5,000 cm^{-1} and 10,000 cm^{-1} are attributed to spectator sites. The 15,100 cm^{-1} MCD band of α-Fe(II) is sensitive to both field and temperature. This defines α-Fe(II) as paramagnetic at low temperatures. The oxo-bridged[16] and hydroxo-bridged[17] 2Fe(II) structures proposed for α-Fe(II) are therefore excluded, as they would be strongly antiferromagnetically coupled (with an antiferromagnetic coupling constant of more than 10 cm^{-1})[18].

The temperature and field dependence of the 15,100 cm^{-1} MCD feature provides direct insight into the ground state of α-Fe(II)—even in the presence of paramagnetic spectator sites. VTVH-MCD isofields for the 15,100 cm^{-1} band are shown in Fig. 2b. The field dependence of the low-temperature saturation limit reflects variation of the α-Fe(II) ground state with field, indicating that α-Fe(II) is a non-Kramers (integer-spin) system[11,18]. The influence of temperature and field on MCD intensity from a non-Kramers doublet is parameterized by an effective g value (g_{eff}) and a rhombic zero-field splitting (ZFS), δ (Fig. 2c). g_{eff} and δ are related to molecular-spin Hamiltonian parameters, and therefore encode information about site geometry and electronic structure (see Methods for detail). VTVH-MCD isotherms from the 15,100 cm^{-1} band plotted against the ratio of the applied magnetic field and the temperature (H/T) (Fig. 2d) are highly nested (that is, they do not overlap), indicating that δ is large for α-Fe(II)[11,18]. Fitting a non-Kramers-doublet model to VTVH-MCD data enables quantification of δ at 9 cm^{-1} and g_{eff} at 8.6. A g_{eff} close to 8 is consistent with either mononuclear high-spin ($S = 2$) Fe(II), or a high-spin 2Fe(II) dimer with oppositely signed ZFS for each Fe(II)[11,18]. We exclude the latter possibility with Mössbauer data (Extended Data Fig. 3), which show a single quadrupole doublet for α-Fe(II) (isomer shift (IS) = 0.89 mm s^{-1}; quadrupole splitting ($|QS|$) = 0.55 mm s^{-1}; 93% of total Fe). VTVH-MCD therefore defines α-Fe(II) as a mononuclear, square planar Fe(II) site.

Different coordination geometries can be distinguished on the basis of the sign of the axial ZFS parameter, D. A large δ of 9 cm^{-1} is well outside the range possible for a negative ZFS ground state (for which δ is generally smaller than 6 cm^{-1}). D is therefore positive for α-Fe(II) (see Fig. 2e and Methods)[11,18,19]. Fitting a positive ZFS $S = 2$ model to VTVH-MCD data enables quantification of the axial and rhombic ZFS parameters, $D = 13 \pm 1$ cm^{-1} and $E = 1.8 \pm 0.5$ cm^{-1} respectively, as well as a molecular $g_{\perp} = 2.15$ (see Methods for detail). Coupled with electronic structure calculations (*vide infra*), these parameters provide detailed insight into the ligand field of α-Fe(II).

With α-Fe(II) defined as a high-spin monomer, additional information can be extracted from its ligand-field spectrum and spin Hamiltonian parameters. The high energy (>12,000 cm^{-1}) of the 15,900 cm^{-1} DR-UV-vis ligand-field band rules out octahedral, tetrahedral and trigonal bipyramidal geometry for α-Fe(II). Square pyramidal (weak axial) geometry leads to a negative ZFS ground state, and is excluded[11]. Square planar structures, however, are associated with spin Hamiltonian and Mössbauer parameters that are highly similar to those of α-Fe(II)[13–15,20,21]. On this basis, we assign α-Fe(II) as a square planar $S = 2$ Fe(II) site. The ligand-field origins of the spectroscopic features of α-Fe(II) are presented in the Methods.

Within the BEA lattice, only the β-type six-membered ring (β-6MR) motifs shown in Fig. 3a have the appropriate geometry to stabilize a CH_4-accessible square planar site[22]. These three β-6MRs yield highly similar Fe(II)-bound sites (Extended Data Fig. 4); ring A1 is presented here. Data presented in Extended Data Fig. 5 (see Methods for detail) indicate that α-Fe(II) only forms in β-6MRs containing two aluminium T-sites (anionic AlO_4 tetrahedra), and that three configurations of these aluminium T-sites are possible for BEAs that have a silicon/aluminium ratio of more than 10 (a silicon/aluminium ratio of more than 12 was used here)[23,24]. We created density functional theory (DFT) cluster models to evaluate the influence of aluminium configuration on the resulting Fe(II)-bound site (Fig. 3b). The cluster models were geometry optimized on the quintet surface, and are all approximately square planar. T4/T4′ and T8/T8′ bind Fe(II) with two neutral $_{\text{Si}}O_{\text{Si}}$ ligands and two



	Experiment	T4/T4′	T6/T6′	T8/T8′
$h\nu(z^2\rightarrow x^2-y^2)$ (cm^{-1})	15,900	14,540	16,750	14,260
D (cm^{-1})	+13 ± 1	–28.4	+13.5	–14.5
E/D	0.14 ± 0.04	0.129	0.078	0.282
$g_\perp$	2.15	1.97(x), 2.08(y)	2.22(x), 2.17(y)	2.00(x), 2.12(y)
Binding energy (kcal mol^{-1})	—	–588	–617	–592
d(Fe-O) (Å)	—	1.95, 2.18	1.99, 2.01	1.98, 2.11

Figure 3 | Computational evaluation of α-Fe(II) cluster models. **a**, β-6MR motifs found in BEA polymorphs A (top) and B (bottom). **b**, Top, cluster models of ring A1 containing distinct configurations of Al tetrahedral (T)-sites (in blue; other atoms omitted for clarity); bottom, CASPT2-predicted spectral features of these Fe(II)-bound sites ($h\nu$—energy of the highest ligand field transition, D, E/D and $g_\perp$), and DFT-predicted Fe(II)-binding energies and iron-to-oxygen bond lengths (d(Fe–O)). Experimental data are provided for comparison in the first column of the table.

anionic $_{Si}O_{Al}$ ligands, while T6/T6′ binds Fe(II) with four anionic $_{Si}O_{Al}$ ligands.

We calculated spectroscopic features of the DFT-optimized cluster models at the CASPT2 level of theory (Fig. 3b and Extended Data Table 1; CASPT2 is the second-order complete active space perturbation theory). CASPT2 predicts a single, high-energy ligand-field band for all three models, thus reproducing the ligand-field spectrum of α-Fe(II) that we defined experimentally. The T4/T4′ and T8/T8′ models fail to reproduce the positive ZFS ground state of α-Fe(II), however (see Extended Data Fig. 6 and Methods). But T6/T6′ accurately reproduces both the spin Hamiltonian parameters and the ligand-field spectrum of α-Fe(II). We therefore assign α-Fe(II) as a high-spin, square planar Fe(II) site bound to a β-6MR by four anionic $_{Si}O_{Al}$ ligands. DFT-calculated Mössbauer parameters of the cluster models further support this assignment (see Methods for detail).

α-O is generated by transferring an O atom from N_2O to α-Fe(II). Low-temperature MCD data from N_2O-activated Fe(II)-BEA are shown in Fig. 4a, along with DR-UV-vis data for comparison. At least five features are observed: three relatively sharp negative bands between 5,000 cm^{-1} and 8,000 cm^{-1}, and a prominent positive band at 20,100 cm^{-1} with a low-energy shoulder. These features are not present before N_2O activation and decay upon reaction with CH_4 (see Fig. 2a and Extended Data Fig. 7), and are therefore new spectroscopic handles of α-O. We collected VTVH-MCD data from the 20,100 cm^{-1} feature (Fig. 4a, bottom). Field dependence is observed in the low-temperature saturation limit of the VTVH-MCD isofields, indicating that α-O—like α-Fe(II)—has an integer-spin ground state. Fitting a non-Kramers-doublet model to VTVH-MCD data yields a $g_{\rm eff}$ of 8.0 and a δ of 7 cm^{-1}. A $g_{\rm eff}$ of 8.0 for the ground state of α-O is consistent with either a mononuclear $S=2$ site or with a weakly coupled dimer of $S=2$ sites with oppositely signed ZFS values. The latter is excluded by Mössbauer data, which show a single quadrupole doublet for α-O (IS = 0.30 mm s^{-1}; |QS| = 0.50 mm s^{-1}; Extended Data Fig. 3). α-O is therefore mononuclear, and a large δ of 7 cm^{-1} indicates a positive ZFS (+D) ground state[11,18]. Fitting a positive ZFS $S=2$ model to VTVH-MCD data (see Methods) results in estimates of $D=8\pm1$ cm^{-1} and $E=0.5\pm0.5$ cm^{-1} for α-O.

VTVH-MCD data thus indicate that transferring an O atom to α-Fe(II) yields a mononuclear $S=2$[FeO]$^{2+}$ site. Two electronic structures are possible: Fe(III)–O$^{\bullet-}$ or Fe(IV)=O. A D substantially larger than 2 cm^{-1} argues against an Fe(III)–O$^{\bullet-}$ ground state[18,25]. On the other hand the sign and magnitude of D for α-O are similar to known $S=2$ Fe(IV)=O sites[26,27]. Spin Hamiltonian parameters coupled with electronic structure calculations (*vide infra*) therefore define α-O to be a mononuclear $S=2$ Fe(IV)=O species.

To define factors contributing to the reactivity of α-O, we added an O atom to the cluster model of α-Fe(II). This results in a square pyramidal Fe(IV)=O species with a short, 1.59-Å Fe–O bond (Fig. 4b, top). The weak ligand field provided by the BEA lattice stabilizes the $S=2$ model (α-^{5}Fe(IV)=O), reproducing the electronic structure and Mössbauer parameters of α-O defined experimentally (see above and Methods). Computational evaluation of the reaction of α-O with CH_4 shows that H-atom abstraction (HAA) occurs on the quintet surface with a low activation barrier of 3.6 kcal mol^{-1} (Fig. 4b, middle)—consistent with the room-temperature CH_4 reactivity observed in ref. 1, and paralleling the computational results of ref. 28.

The CH_4 HAA reaction is slightly exothermic, indicating that the O–H bond strength of the Fe(III)–OH first product and the C–H bond strength of CH_4 are similar. This markedly reduces the barrier for HAA. The exceptional strength of the product O–H bond is related to the axial square pyramidal geometry of α-O, which is unstable in the absence of the zeolite lattice (see Extended Data Fig. 8 and Methods)[29,30]. Lattice constraints destabilize the ^{5}Fe(IV)=O site by about 6 kcal mol^{-1}, increasing the driving force for O–H bond formation. Adding a *trans* axial ligand mitigates this effect. Spin pairing also weakens the O–H bond (by about 12 kcal mol^{-1}; the difference between the reaction enthalpies on the quintet versus triplet surfaces; see Fig. 4b).

Given that the enthalpies of the reactant C–H bond and the product O–H bond are similar, a small $\Delta H^{\ddagger}$ (enthalpy of activation) indicates that α-O has high intrinsic reactivity towards HAA. At the transition-state geometry (when the Fe–O bond elongates from 1.59 Å to 1.72 Å), the Fe(IV)=O unit gains notable radical character, corresponding to a highly reactive Fe(III)–O$^{\bullet-}$ species. This contribution to reactivity also derives from the vacant *trans* axial position, which stabilizes the Fe(3d_{z^2}) orbital, resulting in a highly covalent Fe=O unit. Thus, the high Fe/O covalency and the exchange stabilization of the Fe(3d) manifold lead to an Fe(IV)=O bond that strongly spin-polarizes towards Fe(III)–O$^{\bullet-}$ at the transition-state geometry.

This direct experimental insight into the active site of low-temperature methane hydroxylation in Fe-zeolites establishes VTVH-MCD as a

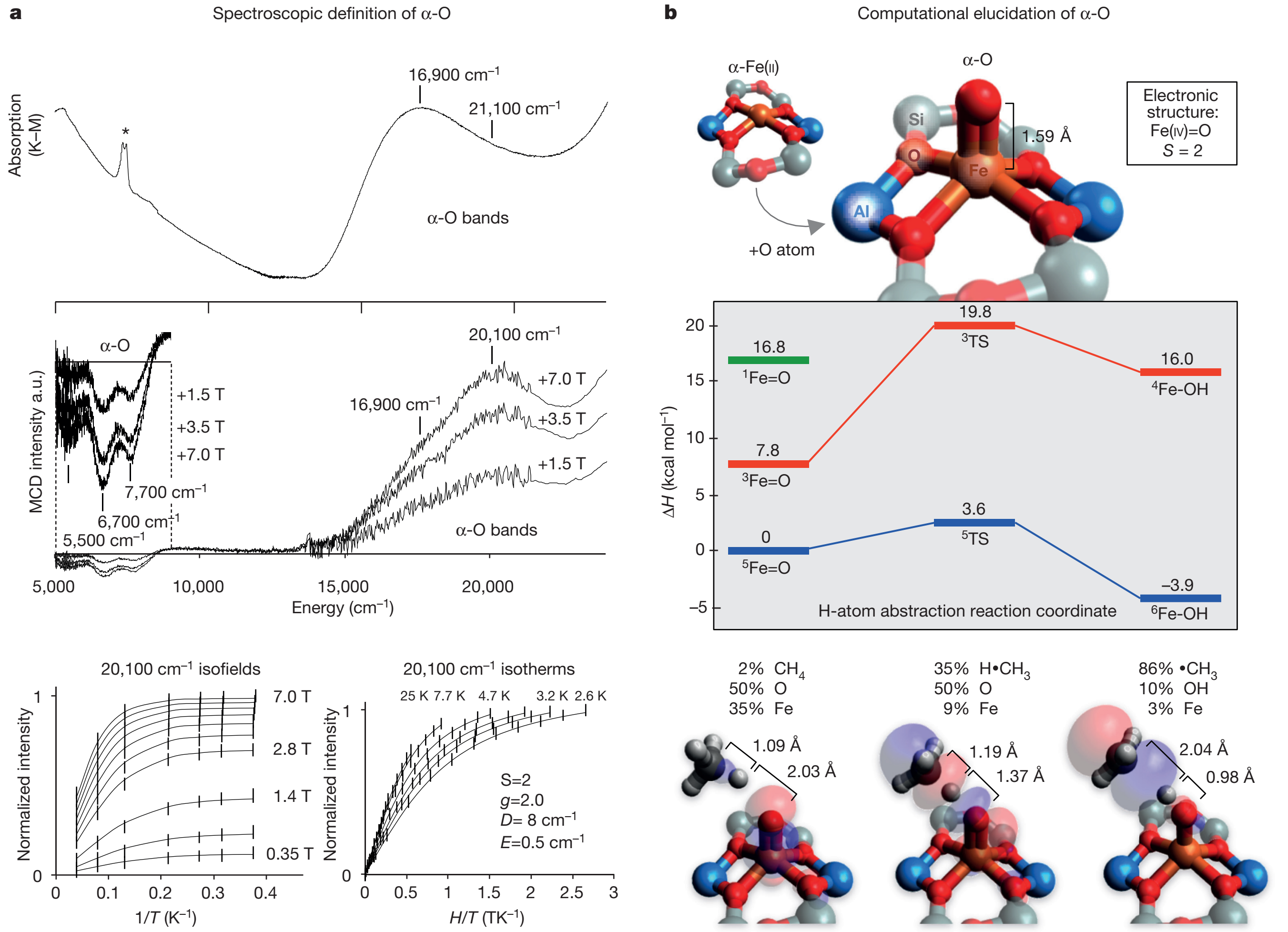


Figure 4 | Spectroscopic and computational elucidation of α-O. a, Top, room-temperature DR-UV-vis data (* = OH overtone), and middle, 3 K MCD data from N_2O-activated BEA. Bottom, VTVH-MCD saturation magnetization data from the 20,100 cm^{-1} band of α-O, including $\pm 1\sigma$ error bars and fit (black curves) to a positive ZFS $S = 2$ model (see Fig. 2e, right), with parameters given in inset. **b**, Top, DFT-optimized structure of α-^{5}Fe(IV)=O in the $S = 2$ ground state. Middle, energetics of the CH_4 HAA reaction. Bottom, evolution of the lowest unoccupied molecular orbital along the reaction coordinate (reactant, left; transition state, middle; and product, right).

powerful, site-selective probe of metal active sites and reactive intermediates in heterogeneous systems, delivering the level of spectroscopic insight into heterogeneous catalysis that has long been available in bioinorganic chemistry. By defining the geometric and electronic structures of α-Fe(II) and α-O and determining how these correlate to reactivity, we have been able to show that the vacant *trans* axial position of α-O, which is enforced by constraints from the zeolite lattice, is a key determinant of the intermediate's exceptional reactivity with CH_4. Thus, by preventing geometric distortion, the zeolite lattice activates α-O for reaction with CH_4, cleaving the strongest aliphatic C–H bond at room temperature to form methanol. In biology, an analogous state in which rigid constraints are used to tune enzyme metal sites for function is known as the 'entatic' state[6], and it will be important to explore further how this can be used to tune the reactivity of metal sites in heterogeneous systems.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 24 March; accepted 13 June 2016.

1. Dubkov, K. A. *et al.* Kinetic isotope effects and mechanism of biomimetic oxidation of methane and benzene on FeZSM-5 zeolite. *J. Mol. Catal. A.* **123,** 155–161 (1997).
2. Dubkov, K. A., Sobolev, V. I. & Panov, G. I. Low-temperature oxidation of methane to methanol on FeZSM-5 zeolite. *Kinet. Catal.* **39,** 72–79 (1998).
3. Ovanesyan, N. S., Shteinman, A. A., Dubkov, K. A., Sobolev, V. I. & Panov, G. I. The state of iron in the Fe-ZSM-5–N_2O system for selective oxidation of methane to methanol from data of Mössbauer spectroscopy. *Kinet. Catal.* **39,** 792–797 (1998).
4. Panov, G. I. Advances in oxidation catalysis, oxidation of benzene to phenol by nitrous oxide. *CATTech* **4,** 18–31 (2000).
5. Smeets, P. J., Woertink, J. S., Sels, B. F., Solomon, E. I. & Schoonheydt, R. A. Transition-metal ions in zeolites: coordination and activation of oxygen. *Inorg. Chem.* **49,** 3573–3583 (2010).
6. Vallee, B. L. & Williams, R. Metalloenzymes: the entatic nature of their active sites. *Proc. Natl Acad. Sci. USA* **59,** 498–505 (1968).
7. Zecchina, A., Rivallan, M., Berlier, G., Lamberti, C. & Ricchiardi, G. Structure and nuclearity of active sites in Fe-zeolites: comparison with iron sites in enzymes and homogeneous catalysts. *Phys. Chem. Chem. Phys.* **9,** 3483–3499 (2007).
8. Centi, G., Genovese, C., Giordano, G., Katovic, A. & Perathoner, S. Performance of Fe-BEA catalysts for the selective hydroxylation of benzene with N_2O. *Catal. Today* **91–92,** 17–26 (2004).
9. Jíša, K. *et al.* Role of the Fe-zeolite structure and iron state in the N_2O decomposition: Comparison of Fe-FER, Fe-BEA, and Fe-MFI catalysts. *J. Catal.* **262,** 27–34 (2009).
10. Baerlocher, C., McCusker, L. B. & Olson, D. H. *Atlas of Zeolite Framework Types* (Elsevier, Amsterdam, 2007).
11. Solomon, E. I., Pavel, E. G., Loeb, K. E. & Campochiaro, C. Magnetic circular dichroism spectroscopy as a probe of the geometric and electronic structure of non-heme ferrous enzymes. *Coord. Chem. Rev.* **144,** 369–460 (1995).
12. Burns, R. G., Clark, M. G. & Stone, A. J. Vibronic polarization in the electronic spectra of gillespite, a mineral containing iron (II) in square-planar coordination. *Inorg. Chem.* **5,** 1268–1272 (1966).
13. Cantalupo, S. A., Fiedler, S. R., Shores, M. P., Rheingold, A. L. & Doerrer, L. H. High-spin square-planar Co(II) and Fe(II) complexes and reasons for their electronic structure. *Angew. Chem. Int. Edn* **51,** 1000–1005 (2012).

14. Pinkert, D. *et al.* A dinuclear molecular iron (II) silicate with two high-spin square-planar FeO_4 units. *Angew. Chem. Int. Edn* **52,** 5155–5158 (2013).
15. Pascualini, M. E. *et al.* A high-spin square-planar Fe (II) complex stabilized by a trianionic pincer-type ligand and conclusive evidence for retention of geometry and spin state in solution. *Chem. Sci.* **6,** 608–612 (2015).
16. Xia, H. *et al.* Direct spectroscopic observation of Fe (III)-phenolate complex formed from the reaction of benzene with peroxide species on Fe/ZSM-5 at room temperature. *J. Phys. Chem. C* **112,** 9001–9005 (2008).
17. Dubkov, K. A., Ovanesyan, N. S., Shteinman, A. A., Starokon, E. V. & Panov, G. I. Evolution of iron states and formation of α-sites upon activation of FeZSM-5 zeolites. *J. Catal.* **207,** 341–352 (2002).
18. Solomon, E. I. *et al.* Geometric and electronic structure/function correlations in non-heme iron enzymes. *Chem. Rev.* **100,** 235–350 (2000).
19. Campochiaro, C., Pavel, E. G. & Solomon, E. I. Saturation magnetization magnetic circular dichroism spectroscopy of systems with positive zero-field splittings: application to $FeSiF_6{\bullet}6H_2O$. *Inorg. Chem.* **34,** 4669–4675 (1995).
20. Clark, M. G., Bancroft, G. M. & Stone, A. J. Mössbauer spectrum of Fe^{2+} in a square-planar environment. *J. Chem. Phys.* **47,** 4250–4261 (1967).
21. Wurzenberger, X., Piotrowski, H. & Klüfers, P. A stable molecular entity derived from rare iron (II) minerals: the square-planar high-spin-d6 $FeIIO_4$ chromophore. *Angew. Chem. Int. Edn* **50,** 4974–4978 (2011).
22. Newsam, J. M., Treacy, M. M. J., Koetsier, W. T. & de Gruyter, C. B. Structural characterization of zeolite beta. *Proc. R. Soc. Lond. A* **420,** 375–405 (1988).
23. Dědeček, J., Sobalík, Z. & Wichterlová, B. Siting and distribution of framework aluminium atoms in silicon-rich zeolites and impact on catalysis. *Catal. Rev. Sci. Eng.* **54,** 135–223 (2012).
24. Lowenstein, W. The distribution of aluminum in the tetrahedra of silicates and aluminates. *Am. Mineral.* **39,** 92–96 (1954).
25. Kahn, O. *Molecular Magnetism* (VCH, New York, 1993).
26. Puri, M. & Que, L. Toward the synthesis of more reactive S=2 non-heme oxoiron (IV) complexes. *Acc. Chem. Res.* **48,** 2443–2452 (2015).
27. McDonald, A. R. & Que, L. High-valent nonheme iron-oxo complexes: synthesis, structure, and spectroscopy. *Coord. Chem. Rev.* **257,** 414–428 (2013).
28. Rosa, A., Ricciardi, G. & Baerends, E. J. Is $[FeO]^{2+}$ the active center also in iron containing zeolites? A density functional theory study of methane hydroxylation catalysis by Fe-ZSM-5 zeolite. *Inorg. Chem.* **49,** 3866–3880 (2010).
29. Rossi, A. R. & Hoffman, R. Transition metal pentacoordination. *Inorg. Chem.* **14,** 365–374 (1975).
30. Neidig, M. L. *et al.* Spectroscopic and electronic structure studies of aromatic electrophilic attack and hydrogen-atom abstraction by non-heme iron enzymes. *Proc. Natl Acad. Sci. USA* **103,** 12966–12973 (2006).
31. Weckhuysen, B. M. & Schoonheydt, R. A. Recent progress in diffuse reflectance spectroscopy of supported metal oxide catalysts. *Catalysis Today* **49,** 441–451 (1999).

Supplementary Information is available in the online version of the paper.

Acknowledgements B.E.R.S. acknowledges support from the National Science Foundation Graduate Research Fellowship Program under grant DGE-11474, and from the Munger, Pollock, Reynolds, Robinson, Smith & Yoedicke Stanford Graduate Fellowship. P.V. acknowledges Research Foundation–Flanders (FWO; grant 12L0715N) and KU Leuven for his postdoctoral fellowships and travel grants during his stay at Stanford University. S.D.H. acknowledges FWO for a PhD (aspirant) Fellowship. L.U. acknowledges FWO for a postdoctoral fellowship. Funding for this work was provided by the National Science Foundation (grant CHE-1360046 to E.I.S.), and within the framework of FWO (grants G0A2216N to B.F.S and G.0865.13 to K.P.). The computational resources and services used for the CASPT2 calculations were provided by the VSC (Flemish Supercomputer Center) and funded by the Hercules Foundation and the Flemish Government department EWI.

Author Contributions E.I.S., B.F.S., R.A.S. and K.P. designed the experiments. B.E.R.S., P.V., M.L.B. and L.H.B. performed the experiments. B.E.R.S. performed the DFT calculations with help from L.H.B. S.D.H. and L.U. performed the CASPT2 calculations. B.E.R.S., P.V. and E.I.S. analysed the data. B.E.R.S. and E.I.S. wrote the manuscript with help from P.V., S.D.H. and L.U.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to E.I.S. (edward.solomon@stanford.edu), B.F.S. (bert.sels@biw.kuleuven.be) or R.A.S. (robert.schoonheydt@biw.kuleuven.be).

Reviewer Information *Nature* thanks A. Bell, E. Bill and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Preparation of Fe-zeolites. We prepared samples of Fe-zeolites (Fe-beta (BEA), Fe-ZSM-5 (having a MFI topology), Fe-ferrierite (FER) and Fe-mordenite (MOR)) from the corresponding acid zeolites by diffusion impregnation of Fe(acac)$_3$ (acac = acetylacetonate) in toluene solution (10 mM; 25 ml g^{-1} zeolite). The Fe(acac)$_3$ concentration was lowered to 2.5 mM when preparing Fe-BEA samples with a lower iron content (0.3 wt%). ^{57}Fe-enriched Fe(II) BEA was synthesized similarly using a 50% mixture of labelled and non-labelled Fe(acac)$_3$. All samples were calcined in air at 550 °C to remove organic material. This preparation method minimizes Fe heterogeneity and limits the formation of oxide/hydroxide species (relative to aqueous exchange or sublimation). The samples, each approximately 1 gram, were then subjected to high-temperature treatment at 900 °C in helium for 2 hours, followed by a reductive treatment in hydrogen at 700 °C for 1 hour. A flow rate of 50 cm^3 min^{-1} was used in both gas treatments. The Fe content of the resulting Fe(II)-BEA was determined by inductively coupled plasma mass spectrometry. All subsequent manipulations were carried out under an inert atmosphere (either nitrogen or helium). N$_2$O activation of Fe(II)-BEA (1 gram) was performed at 250 °C using a 5 vol% N$_2$O/helium flow (50 cm^3 min^{-1}). From Mössbauer (Extended Data Fig. 3), a typical sample of N$_2$O-activated Fe-BEA (0.3 wt% Fe, Si/Al = 12) contains 42 μmol α-O per gram of catalyst (78% of total Fe content). The CH$_4$ reaction was performed at room temperature using a 10 vol% CH$_4$/helium flow (50 cm^3 min^{-1}). We recovered 30–35 μmol methanol per gram of catalyst through liquid extraction, corresponding to a 70–80% methanol yield (some methanol remains adsorbed on the zeolite).

Diffuse reflectance spectroscopy. DR-UV-vis spectra were recorded on a Varian Cary 5000 UV–VIS–NIR spectrophotometer at room temperature against a halon white reflectance standard in the range 4,000–50,000 cm^{-1}. All treatments before *in situ* UV-vis-NIR spectroscopic measurements were performed in a quartz U-tube/flow cell. The latter was equipped with a window for *in situ* UV-vis-NIR diffuse reflectance spectroscopy. After UV-vis measurement, the catalyst pellets were mounted in the quartz side arm of the U-tube, ensuring that the conditions for UV-vis-NIR and MCD/Mössbauer measurements were identical.

MCD and VTVH-MCD spectroscopy. All cells for MCD spectroscopy were prepared under an inert nitrogen atmosphere. Mull samples with suitable transparency were prepared using dried and degassed perfluoro-2-methylpentane as the mulling agent. This inert glassing agent can be rigorously dried and degassed, preserving the oxidation state and coordination environment of Fe sites in Fe-BEA. Further, its refractive index is a suitable match to Al/SiO$_2$ materials, minimizing signal attenuation from scattering. Mulls were frozen using liquid nitrogen immediately after preparation, with special care taken to avoid exposure to oyxgen. MCD data were collected using a Jasco J730 spectropolarimeter with a liquid nitrogen cooled InSb detector coupled to an Oxford Instruments SM4000-7T superconducting magnet. MCD spectra were background corrected for baseline effects using data collected at 0 T. VTVH-MCD data were collected using a calibrated Cernox temperature sensor (Lakeshore Cryotronics, calibrated to 1.5–300 K with 0.001 K tolerance) inserted directly into the sample cell. Isotherms/isofields were normalized to the maximum observed intensity. Spin Hamiltonian parameters were extracted from VTVH-MCD isotherms using established procedures (see below)[11,18,19,32].

VTVH-MCD analysis. As shown in Fig. 2c, rhombic ZFS mixes the ±M_s levels of a non-Kramers doublet into equal admixtures designated $|X\rangle$ and $|Y\rangle$, split by δ. $|X\rangle$ and $|Y\rangle$ are both MCD silent owing to the cancellation of oppositely signed ±M_s contributions to MCD intensity. Magnetic field parallel to the principal axis of the ZFS tensor induces MCD activity in a non-Kramers doublet by mixing $|X\rangle$ and $|Y\rangle$ through an off-diagonal Zeeman matrix element. The resulting states are eigenfunctions of the spin Hamiltonian matrix for an isolated non-Kramers doublet:

$$\hat{H} = \begin{array}{c|cc} & |X\rangle & |Y\rangle \\ \hline |X\rangle & +\delta/2 & g_{\text{eff}}\beta H\cos\theta \\ |Y\rangle & g_{\text{eff}}\beta H\cos\theta & -\delta/2 \end{array}$$

In the case that $g_{\text{eff}}\beta H$ (where β is the Bohr magneton and H is the magnetic field) is large relative to δ, pure ±M_s states are recovered, and the $1/T$ saturation limit attains a maximum. The sensitivity of the saturation limit to field in Fig. 2b therefore reflects the magnitude of δ. As shown in Fig. 2e, the information content of δ and g_{eff} for mononuclear high spin Fe(II) depends on the sign of the axial ZFS parameter D. For negative ZFS systems, δ is the rhombic splitting of the $M_s = \pm 2$ doublet, which should be <6 cm^{-1}. The $\delta = 9$ cm^{-1} of α-Fe(II) is well outside this range. For positive ZFS systems, the $M_s = 0$ ground state and one component of the $M_s = \pm 1$ excited state together behave as an effective $M_s = \pm 2$ non-Kramers doublet, with the magnetic field oriented perpendicular to the principal axis of the ZFS tensor. For a positive ZFS ground state, δ is this splitting, which can be large. A δ of 9 cm^{-1} therefore defines α-Fe(II) to be a $+D$ system.

The axial and rhombic ZFS parameters, D and E, can be quantified from VTVH-MCD data by including higher-temperature (4.7–20 K) isotherms. At higher temperatures, the second component of the $M_s = \pm 1$ excited state not included in the non-Kramers doublet model becomes thermally populated (Fig. 2e, right). A three-level model incorporating this component provides estimates of $D = 13 \pm 1$ cm^{-1} and $E = 1.8 \pm 0.5$ cm^{-1}. A perpendicular field is required to induce MCD activity in the $M_s = 0$ ground state of a +ZFS $S = 2$ system (this level is not sensitive to the parallel field). Thus at low temperatures (where $M_s = 0$ alone is populated), the g_{eff} of a +ZFS $S = 2$ system correlates to a molecular g value ($= g_{\text{eff}}/4$)[11,29].

Fitting an $S = 2$ spin Hamiltonian to VTVH-MCD data results in similar estimates of D and E/D (ref. 32). This fitting procedure involves a considerably larger parameter space (three g values, three transition moments, E, and E/D). Initially, all three g values were fixed at the spin-only value of 2.0023 while D and E/D were incremented over the range $-20 \leq D \leq +20$ and $0 \leq E/D \leq 0.3$. This provided estimates of $D = +12$–18 cm^{-1} and $E/D = 0.15$–0.25. Fixing $g_{x,y} = 2.16$ (the value of g being defined from the non-Kramers doublet fit) and $g_z = 2.0023$ resulted in improved fits, but ultimately the same estimates of D (12–18 cm^{-1}) and E/D (0.15–0.25).

Mössbauer spectroscopy. ^{57}Fe Mössbauer spectra were recorded with a See Co. W302 resonant gamma ray spectrometer in horizontal geometry with zero external field using a 1.85 GBq source (Be window, Rh matrix). Data were collected from samples enriched with 50% ^{57}Fe. All spectra were recorded at room temperature and isomer shifts are given relative to α-iron foil at room temperature. Spectra were collected with 1,024 points and summed up to 512 points before analysing, and then fit with Lorentzian doublets using the Vinda software package for Microsoft Excel.

DFT calculations. Cluster models were generated from crystallographic coordinates of BEA polymorphs A and B[22], and dangling O groups were capped with H (see Supplementary Tables 1–3 for coordinates). Spin-unrestricted DFT calculations were performed with Gaussian 09 (ref. 33). The B3LYP functional was used for all DFT calculations. The 6-311G* basis set was used for Fe, for atoms directly coordinated to Fe, and for CH$_4$. The 6-31G* basis set was used for all other atoms. For geometry optimizations, the six T-sites of the six-membered rings were allowed to relax, and all other atoms were frozen (O and Si atoms at their crystallographic positions). All barriers for the CH$_4$ reaction were zero-point corrected. Coordinates of the DFT models of α-Fe(II) and α-Fe(IV)=O are reported in the Supplementary Information, along with coordinates of the α-Fe(IV)=O/CH$_4$ H-atom abstraction transition state. Mössbauer isomer shifts and quadrupole splittings were calculated from the small cluster models used for CASPT2 calculations (see below). Isomer shifts were calculated with the ORCA computational package using the B3LYP functional. The CP(PPP) basis set[34] was used on Fe, with 6-311G* on coordinating O atoms and 6-31G* on all others. A calibration curve was generated by relating the DFT-calculated electron densities at the iron nucleus $|\psi_0|^2$ values to the experimental isomer shifts for a test set of 23 structurally defined Fe complexes. The IS values of the α-Fe models were then estimated from the value of $|\psi_0|^2$ calculated for each cluster model. Quadrupole splittings were calculated using the B3LYP functional, with TZVP on Fe and coordinating O atoms, and 6-31G* on all others.

CASSCF/CASPT2 calculations. To reduce computational time, smaller clusters were constructed out of the DFT-optimized geometries as follows: the six-membered ring containing Fe(II) was cut out of the larger cluster and the dangling O atoms were capped with H at 0.95 Å. *Ab initio* calculations were performed with MOLCAS-8.1 (ref. 35). In the multiconfigurational approach used, relativistic effects were treated in two steps, both based on the Douglas–Kroll Hamiltonian[36]. Scalar terms were included in the basis-set generation and used to determine spin-free wave functions and energies through the use of the complete active space self-consistent field (CASSCF) method. Electron correlation effects were considered by means of the second-order complete active space perturbation theory (CASPT2)[37]. Next, spin-orbit coupling was treated in the mean field (AMFI)[36] by means of the restricted active space state interaction (RASSI) method[38], which uses the optimized CASSCF/CASPT2 wave functions as the basis states. From the resulting spin-orbit eigenstates, we computed the gyromagnetic tensor of the ground state by means of pseudospin $S = 2$ formalism as implemented in the SINGLE_ANISO module, in order to obtain the three main anisotropy axes and the associated gyromagnetic values (g$_x$, g$_y$, and g$_z$)[39]. The spin-orbit interaction was computed within all quintet (5) and all triplet (45) states with the RASSI program. Extended ANO-RCC basis sets[40,41] were used with the following contractions: [7*s*6*p*5*d*3*f*2*g*1*h*] for Fe; [4*s*3*p*2*d*1*f*] for O; [4*s*3*p*1*d*] for Si and Al; and [2*s*1*p*] for H. The active space was chosen according to the standard rules for transition-metal complexes[36]: that is, five 3*d* and five 4*d* orbitals of Fe; and bonding 2*p* orbitals of the coordinating O atoms. The standard iterative phase estimation algorithm shift of 0.25 a.u. (ref. 42) and an imaginary shift of 0.1 a.u. (ref. 43) were used for the zeroth-order Hamiltonian for second-order perturbation theory. Correlation of the core electrons 1*s*, 2*s* and 2*p* of Fe, Si and Al and 1*s* of O was not taken into account in the CASPT2 step.

Spectroscopic features of α-Fe(ii). On the basis of correlation to high-spin square planar Fe(ii) sites and results from CASPT2, we assign the 15,900 cm^{-1} ligand-field band of α-Fe(ii) the $3d_{z^2} \rightarrow 3d_{x^2-y^2}$ transition of a square planar site. The high energy of this transition reflects the equatorial anisotropy of the α-Fe(ii) ligand field, as well as the unique stability of $3d_{z^2}$ in square planar geometry brought about by 4*s* mixing in the absence of axial ligands. The small Mössbauer |QS| of 0.55 mm s^{-1} has similar origins: the combination of an equatorial ligand field with a doubly occupied $3d_{z^2}$ orbital (that is, axial distribution of *d* electron density) leads to near-cancellation of large, oppositely signed lattice and valence contributions (respectively) to QS.

The α-Fe(ii) binding site. As shown in Extended Data Fig. 5, the relative population of α-Fe(ii) is maximized at low Fe loading. Reducing the Fe loading from 1.0 wt% to 0.3 wt% results in a decrease in the population of spectator sites from about 50% to just 7% (see Extended Data Fig. 3). The predominance of α-Fe(ii) at low Fe loadings indicates that α-Fe(ii) is the most stable Fe(ii) site in Fe(ii)-BEA. A sufficiently low lattice Si/Al ratio (that is, a high lattice Al content) must also be attained to stabilize α-Fe(ii). Thus a threshold concentration of lattice Al is required for α-Fe(ii) to form. This indicates that α-Fe(ii) is bound to the BEA lattice by more than one Al T-site. Certain configurations only of aluminium are possible within a β-6MR. Aluminium T-sites separated by a single silicon T-site (that is, Al-O-Si-O-Al sequences) do not exist in BEA with Si/Al ratios greater than 10 (a Si/Al ratio of more than 12 was used here)[23]. Combined with Lowenstein's rule[24], this precludes β-6MRs containing three or more aluminium T-sites. Al-(O-Si)$_2$-O-Al sequences are, however, frequently implicated in the binding of divalent cations to zeolite lattices[23]. The three Al-(O-Si)$_2$-O-Al sequences possible within a β-6MR are shown in Fig. 3b, and these are the only configurations containing multiple Al T-sites that are possible in BEA with Si/Al ratios greater than 10.

Differences in the predicted spin Hamiltonian parameters of the three models shown in Fig. 3—in particular, differences in the sign of *D*—reflect differences in the first coordination shell of Fe. In the case of T6/T6′, Fe(ii) is approximately square planar with four essentially equivalent $_{\mathrm{Si}}\mathrm{O}_{\mathrm{Al}}$ ligands at 1.99–2.01 Å. This is a weak axial system, which leads to a positive *D*. For T4/T4′ and T8/T8′, Fe(ii) is bound by two short (1.95–1.98 Å) Fe–$_{\mathrm{Si}}\mathrm{O}_{\mathrm{Al}}$ bonds and two long (2.11–2.18 Å) Fe–$_{\mathrm{Si}}\mathrm{O}_{\mathrm{Si}}$ bonds. As shown in Extended Data Fig. 5, the $_{\mathrm{Si}}\mathrm{O}_{\mathrm{Al}}$ ligands bind Fe(ii) more strongly, and as a result, the Fe–$_{\mathrm{Si}}\mathrm{O}_{\mathrm{Al}}$ vector (approximately) defines the magnetic *z* axis for these sites. T4/T4′ and T8/T8′ are therefore best described as strong axial systems (with a rhombic perturbation), leading to a negative *D*.

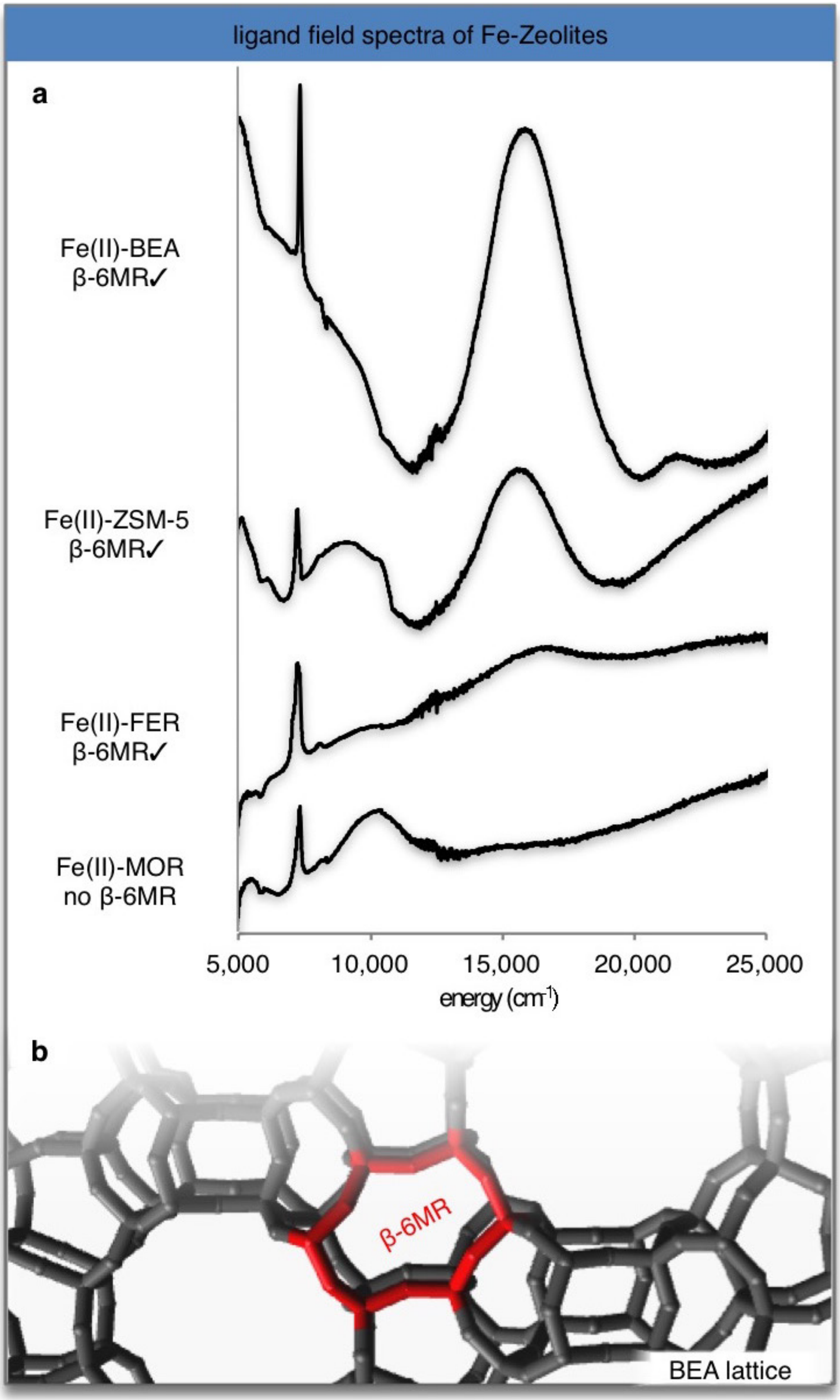
The three cluster models can be further differentiated on the basis of their predicted Mössbauer parameters. The T6/T6′ model reasonably reproduces the low IS and small |QS| values of α-Fe(ii) (calculated IS $=0.72$ mm s^{-1} and QS $=-0.95$ mm s^{-1}; experimental IS $=0.89$ mm s^{-1} and |QS| $=$ 0.55 mm s^{-1}). The T4/T4′ and T8/T8′ Fe(ii)-bound models yield similar calculated values of IS (0.78 mm s^{-1} and 0.81 mm s^{-1}, respectively), but larger values of QS ($+1.38$ mm s^{-1} and $+1.15$ mm s^{-1}, respectively). Reasonable agreement is also achieved for the T6/T6′ model of α-O (calculated IS $=0.22$ mm s^{-1} and QS $=-0.24$ mm s^{-1}; experimental IS $=0.30$ mm s^{-1} and |QS| $=0.50$ mm s^{-1}).

Contributions to the reactivity of α-O. In order to understand the origin of the strong (>100 kcal mol^{-1}) O–H bond of α-^{6}Fe(iii)-OH, we evaluated a set of high-spin Fe(iii)-OH and Fe(iv)=O DFT models (see Extended Data Fig. 8). The models with square pyramidal geometry with an empty coordination site *trans* to the O ligand have the strongest O–H bonds. This geometry destabilizes the ^{5}Fe(iv)=O site relative to the ^{6}Fe(iii)-OH, increasing the driving force for O–H bond formation. This is clearly illustrated with site 1, which features a macrocyclic ligand similar to the β-6MR of a zeolite. Removing an H atom from 1-^{6}Fe(iii)–OH results in 1-^{5}Fe(iv)=O which, like α-^{5}Fe(iv)=O, has a vacant *trans* axial position. The strength of the 1-^{6}Fe(iii)-OH OH bond is calculated to be 101 kcal mol^{-1} (versus the 102 kcal mol^{-1} C–H bond strength calculated for CH$_4$). A second conformation of 1-^{5}Fe(iv)=O is more stable, however. Shifting the oxo from the axial position to the equatorial position results in stronger bonding[29,30]. Exchanging a destabilizing equatorial σ-antibonding interaction (occupied $\sigma^* d_{x^2-y^2}$) for a stabilizing axial σ-bonding interaction (vacant $\sigma^* d_{z^2}$) stabilizes the site by 4.5 kcal mol^{-1}. After correction for the strain in the macrocyclic ligand, this difference increases to 6.0 kcal mol^{-1}. This quantifies the 'entatic' contribution to the strength of the O–H bond in 1-^{6}Fe(iii)-OH.

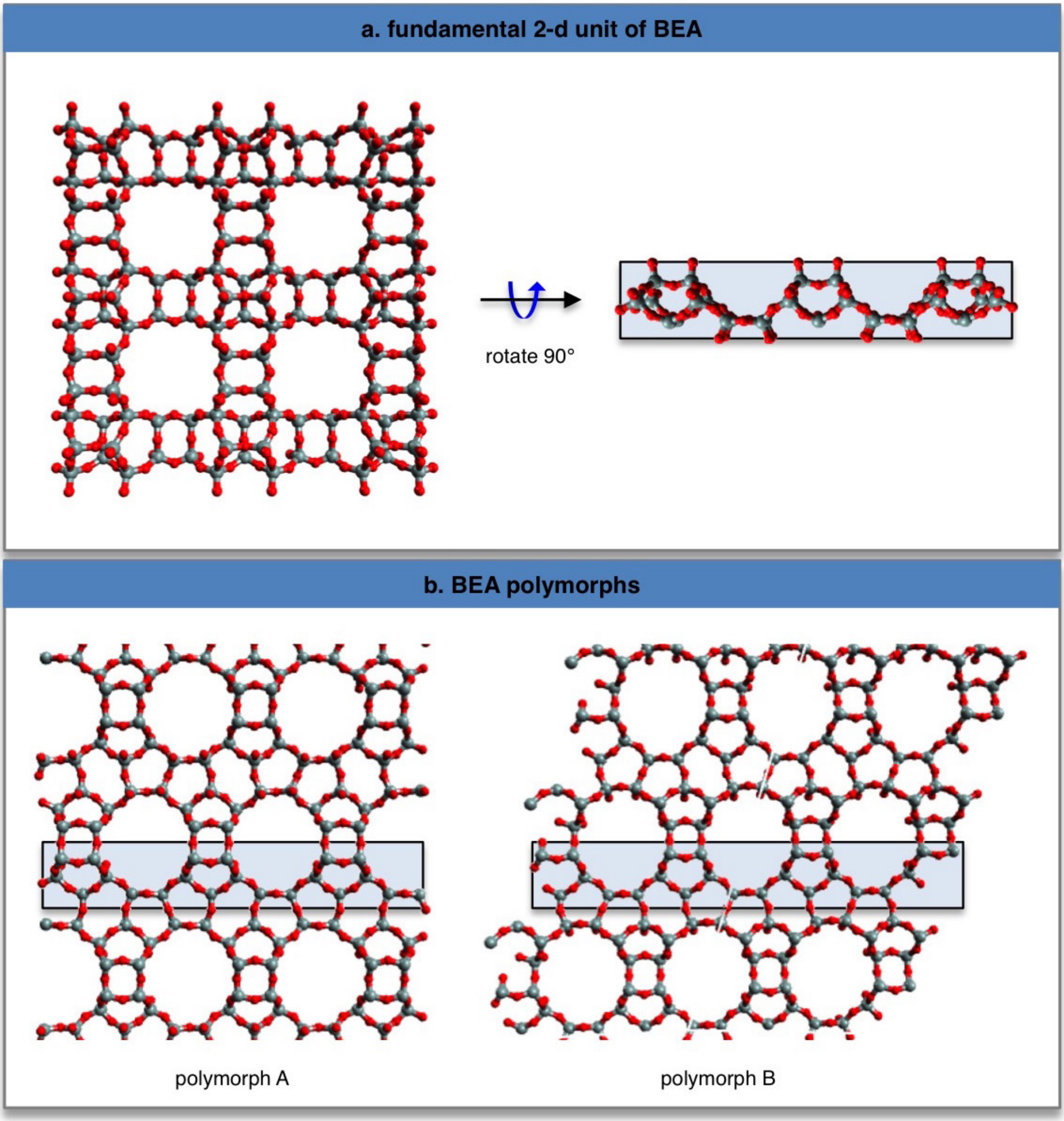
Site 2-^{6}Fe(iii)-OH, which features a similar first coordination sphere to 1-^{6}Fe(iii)-OH but without geometric constraints, has a substantially weaker O–H bond. Removing an H atom from 2-^{6}Fe(iii)-OH causes the O ligand to shift from the axial to the equatorial position. Because of a lack of geometric constraints, the axial isomer is not a stable minimum for 2-^{5}Fe(iv)=O. As a result, there is no entatic contribution to the O–H bond strength for this site, and the O–H bond of 2-^{6}Fe(iii)-OH is weakened by 6 kcal mol^{-1} relative to 1-^{6}Fe(III)-OH (95 kcal mol^{-1} versus 101 kcal mol^{-1}). Constraining site 2 to be square pyramidal destabilizes 2-^{5}Fe(iv)=O site by 6 kcal mol^{-1}, resulting in a similar O–H bond strength to site 1.

Site 3 is qualitatively similar to site 1, in that removing an H atom from 3-^{6}Fe(iii)-OH results in the axial conformation of 3-^{5}Fe(iv)=O. The O–H bond strength of 3-^{6}Fe(iii)-OH is 95 kcal mol^{-1}. Site 4 is formed by adding a *trans* axial ammonia ligand to site 3. The O–H bond of 4-^{6}Fe(iii)-OH is 7 kcal mol^{-1} weaker than that of 3-^{6}Fe(iii)-OH (88 kcal mol^{-1} versus 95 kcal mol^{-1}). Adding an axial ligand mitigates the ligand-field stabilization-energy-based driving force for O–H bond formation (*vide supra*). The additional ligand also stabilizes the oxidized Fe(iv) site over the reduced Fe(iii) site (that is, causing a change in reduction potential), further weakening the O–H bond.

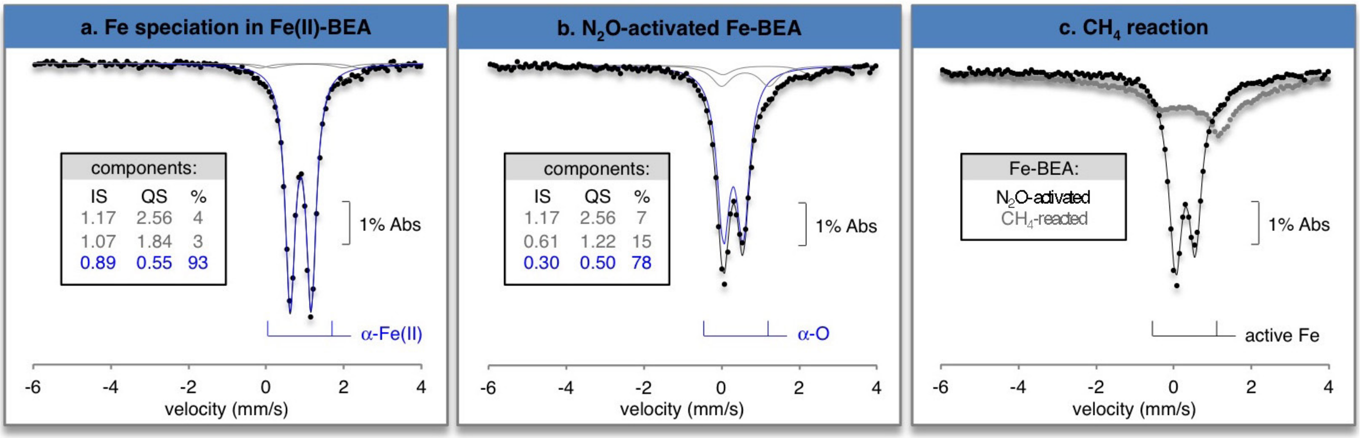
32. Neese, F. & Solomon, E. MCD C-term signs, saturation behavior, and determination of band polarizations in randomly oriented systems with spin *S* $\geq 1/2$. Applications to *S* = 1/2 and *S* = 5/2. *Inorg. Chem.* **38,** 1847–1865 (1999).
33. Frisch, M. J. *et al. Gaussian 09, Revision E.01* (Gaussian, Wallingford, 2009)
34. Neese, F. Prediction and interpretation of the ^{57}Fe isomer shift in Mössbauer spectra by density functional theory. *Inorg. Chim. Acta* **337,** 181–192 (2002).
35. Aquilante, F. *et al.* Molcas 8: new capabilities for multiconfigurational quantum chemical calculations across the periodic table. *J. Comp. Chem.* **37,** 506–541 (2016).
36. Andersson, K., Malmqvist, P.-Å. & Roos, B. O. Second-order perturbation theory with a complete active space self-consistent field reference function. *J. Chem. Phys.* **96,** 1218–1226 (1992).
37. Hess, B. A., Marian, C. M., Wahlgren, U. & Gropen, O. A mean-field spin-orbit method applicable to correlated wavefunctions. *Chem. Phys. Lett.* **251,** 365–371 (1996).
38. Malmqvist, P. A., Roos, B. O. & Schimmelpfennig, B. The restricted active space (ras) state interaction approach with spin–orbit coupling. *Chem. Phys. Lett.* **357,** 230–240 (2002).
39. Chibotaru, L. F. & Ungur, L. Ab initio calculation of anisotropic magnetic properties of complexes. I. Unique definition of pseudospin hamiltonians and their derivation. *J. Chem. Phys.* **137,** 064112 (2012).
40. Roos, B. O., Lindh, R., Malmqvist, P. A., Veryazov, V. & Widmark, P. O. New relativistic ANO basis sets for transition metal atoms. *J. Phys. Chem. A* **109,** 6575–6579 (2005).
41. Roos, B. O., Lindh, R., Malmqvist, P. A., Veryazov, V. & Widmark, P. O. Main group atoms and dimers studied with a new relativistic ano basis set. *J. Phys. Chem. A* **108,** 2851–2858 (2004).
42. Ghigo, B., Roos, B. O. & Malmqvist, P. A. A modified definition of the zeroth-order Hamiltonian in multiconfigurational perturbation theory (CASPT2). *Chem. Phys. Lett.* **396,** 142–149 (2004).
43. Forsberg, N. & Malmqvist, P. A. Multiconfiguration perturbation theory with imaginary level shift. *Chem. Phys. Lett.* **274,** 196–204 (1997).



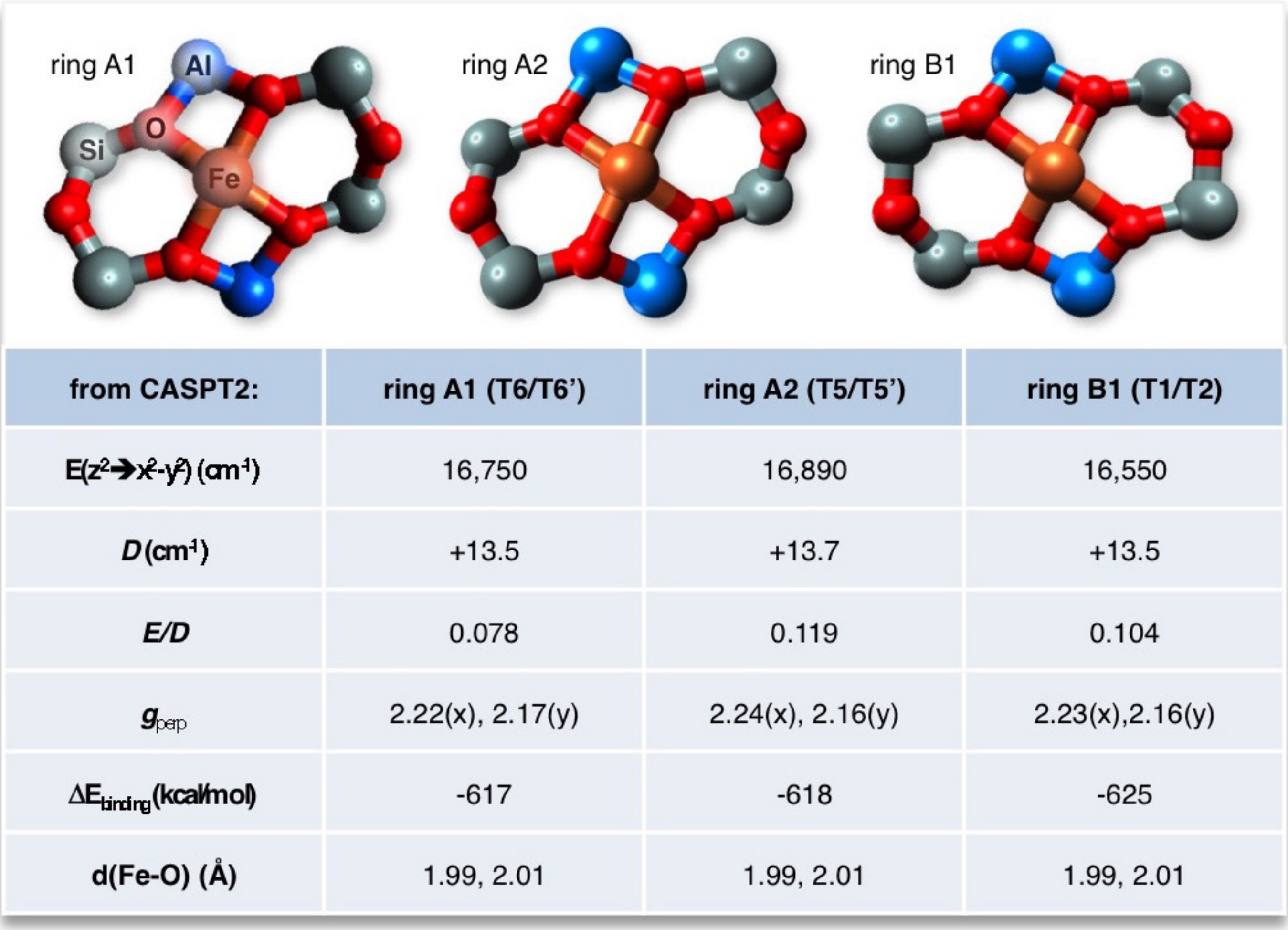
Extended Data Figure 1 | DR-UV-vis spectra of Fe-zeolites. a, Ligand-field DR-UV-vis spectra of three Fe(II)-zeolites that are known to contain α-Fe(II): Fe(II)-BEA, Fe(II)-ZSM-5, and Fe(II)-FER. Fe(II)-MOR, which does not stabilize α-Fe(II), is included for comparison. The lattice topologies that stabilize α-Fe(II) have a conserved structural motif—the β-type six-membered ring (β-6MR). **b**, An example of a β-6MR is highlighted in this BEA lattice[22].



Extended Data Figure 2 | The BEA lattice. a, The structure of the fundamental two-dimensional building unit of BEA. BEA is a layered structure built up from this unit. **b**, BEA is a disordered intergrowth of two polymorphs, BEA-A and BEA-B, which result from different layerings of the same fundamental two-dimensional building unit (highlighted in blue). Both polymorphs feature three-dimensional networks of 10 Å × 10 Å channels, large enough to accommodate CH_4 and other small molecules[22].

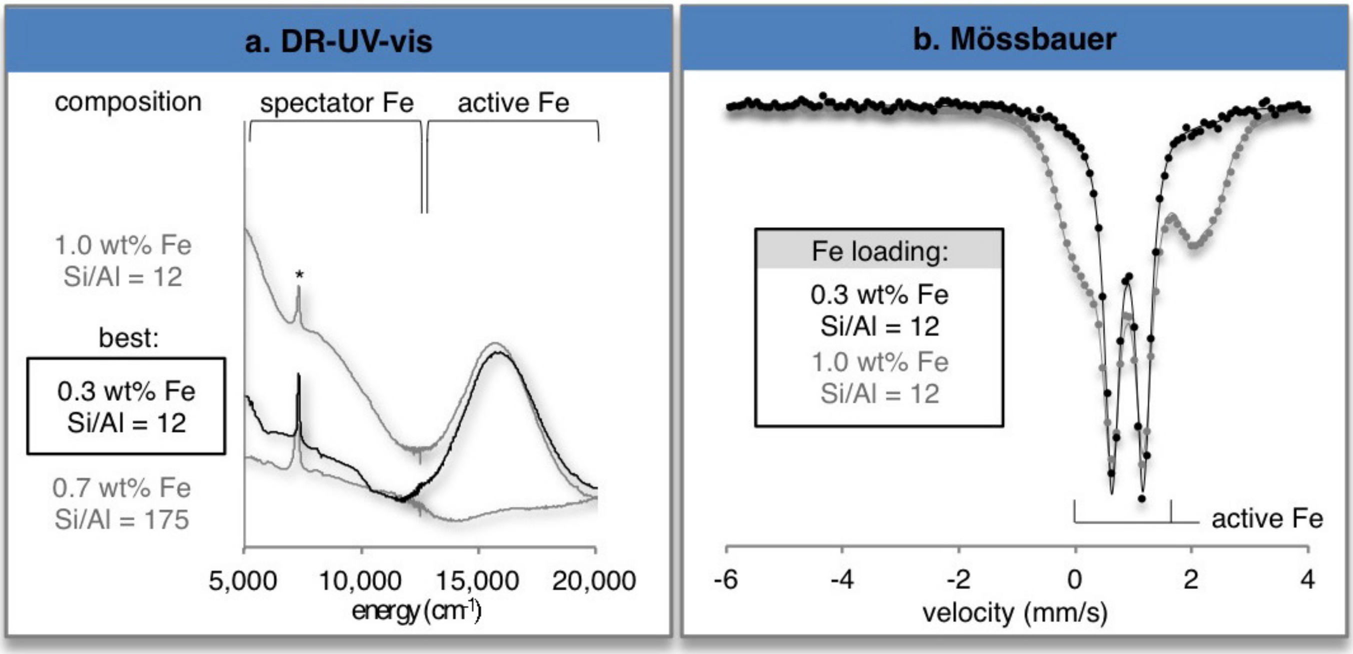


Extended Data Figure 3 | Mössbauer features of Fe-BEA. a, Room-temperature Mössbauer data were collected from a sample of Fe(II)-BEA containing 0.3 wt% Fe. Three Fe components were resolved. Abs, absorption. **b**, Reacting Fe(II)-BEA with N_2O results in loss of the IS $= 0.89$ mm s^{-1} major species and appearance of a new major component (IS $= 0.30$ mm s^{-1}; 78%). **c**, This new major species is eliminated upon reaction with CH_4 at room temperature. It is therefore assigned to α-O. The IS = 0.89 mm s^{-1} component of Fe(II)-BEA is thus assigned to α-Fe(II). Similar Mössbauer features have also been observed in Fe-ZSM-5 and Fe-FER, but they have not been assigned to α-Fe(II)[9].

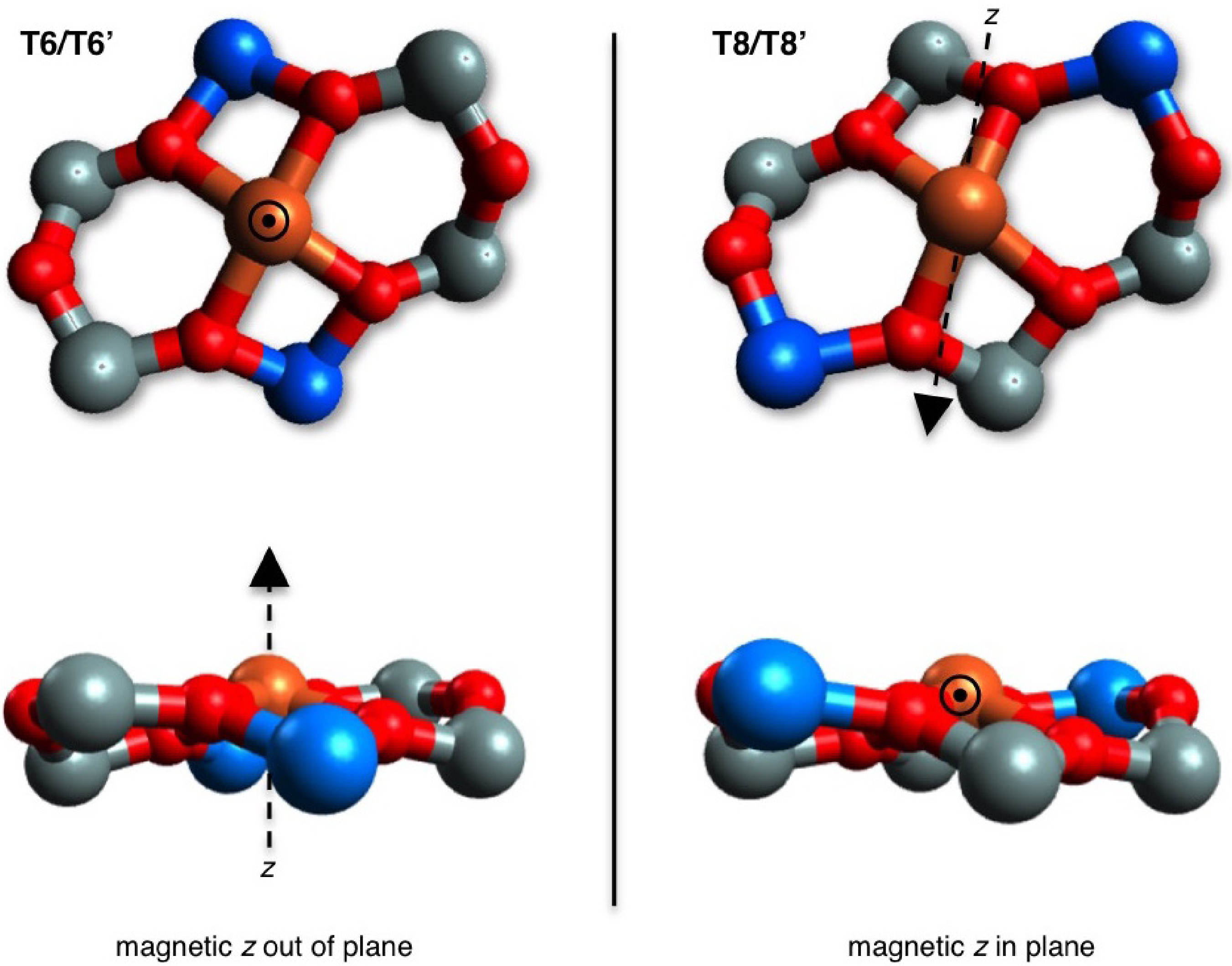


from CASPT2:	ring A1 (T6/T6’)	ring A2 (T5/T5’)	ring B1 (T1/T2)
E(z^2→x^2-y^2) (cm^{-1})	16,750	16,890	16,550
D (cm^{-1})	+13.5	+13.7	+13.5
E/D	0.078	0.119	0.104
g_{perp}	2.22(x), 2.17(y)	2.24(x), 2.16(y)	2.23(x),2.16(y)
$\Delta E_{binding}$ (kcal/mol)	-617	-618	-625
d(Fe-O) (Å)	1.99, 2.01	1.99, 2.01	1.99, 2.01

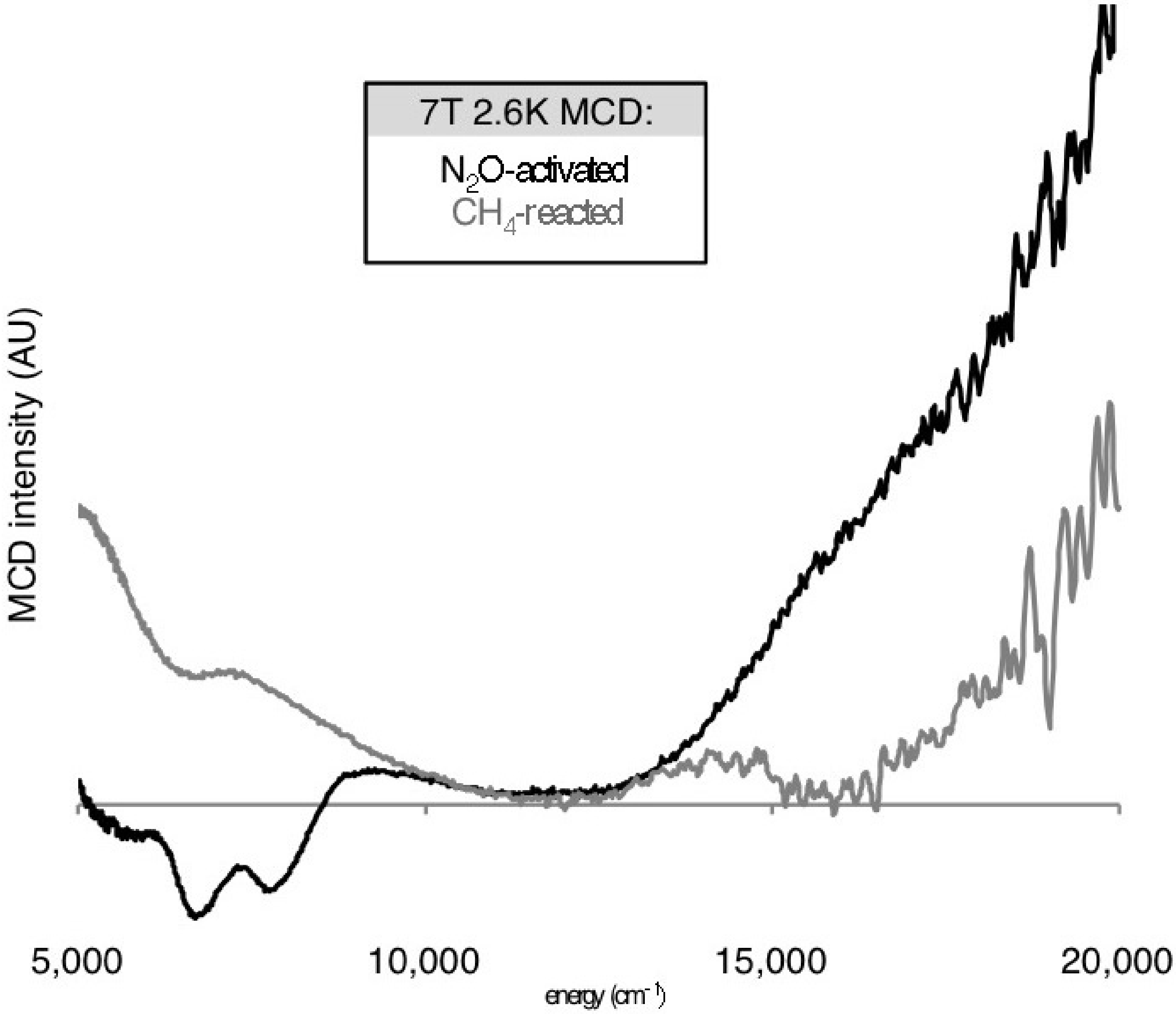
Extended Data Figure 4 | Influence of β-6MR identity on predicted spectral features. DFT-calculated structures of analogous Fe(II) sites formed in each of the three types of β-6MR present in BEA (rings A1 and A2 in polymorph A, and B1 in polymorph B). Other atoms have been omitted for clarity. The table shows that the three sites are highly similar with respect to their metrical parameters, DFT-predicted Fe(II) binding energies, and CASPT2-predicted spectral features.



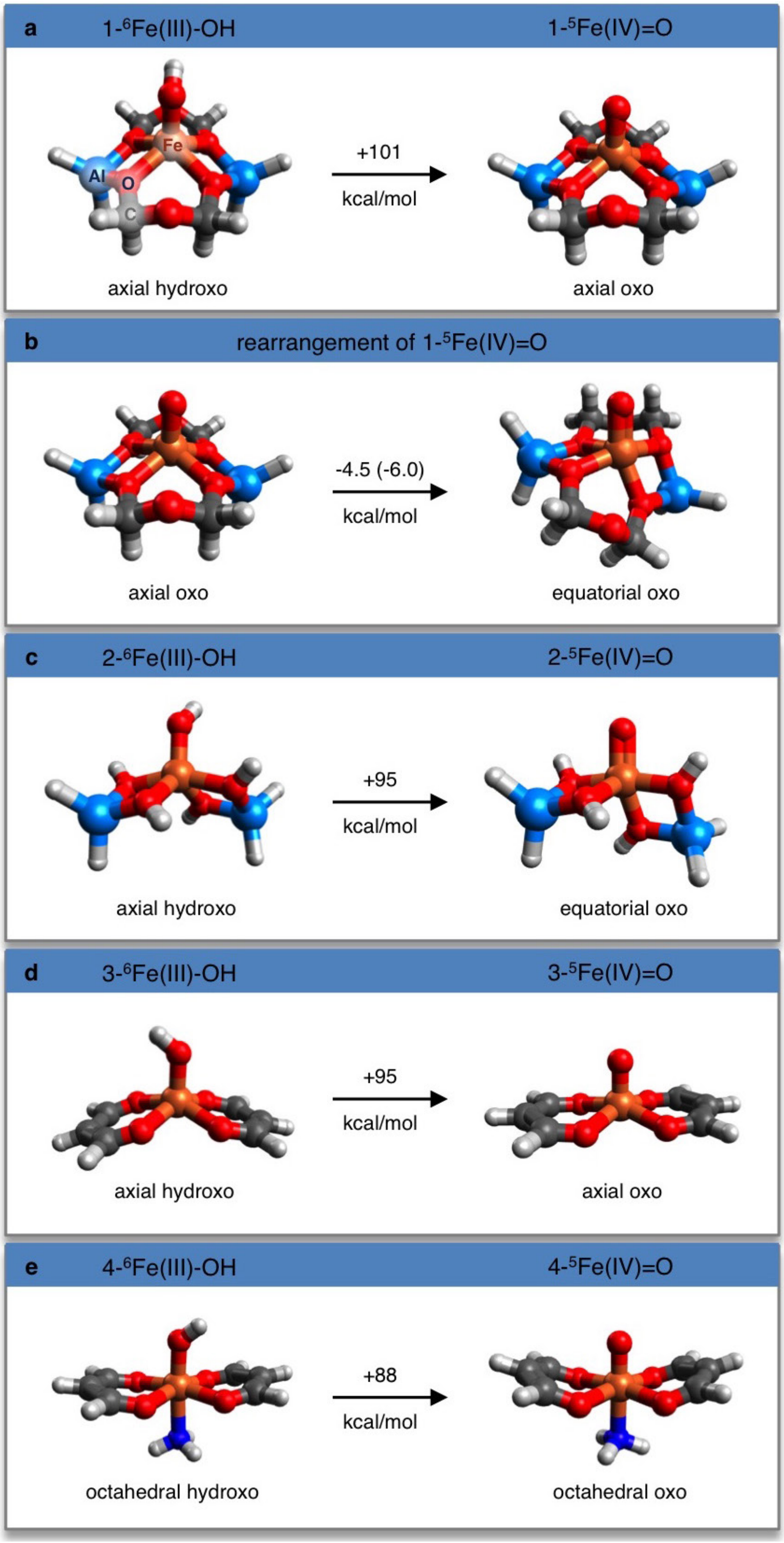
Extended Data Figure 5 | Influence of catalyst preparation on Fe speciation. a, DR-UV-vis spectra (* = OH overtone) and **b**, Mössbauer spectra of Fe(II)-BEA, showing the influence of the lattice Si/Al ratio and Fe loading on Fe speciation.



Extended Data Figure 6 | Magnetic axes of the cluster models. Orientation of the magnetic *z* axes of the T6/T6′ (left) and T8/T8′ (right) cluster models. Atoms have been omitted for clarity.

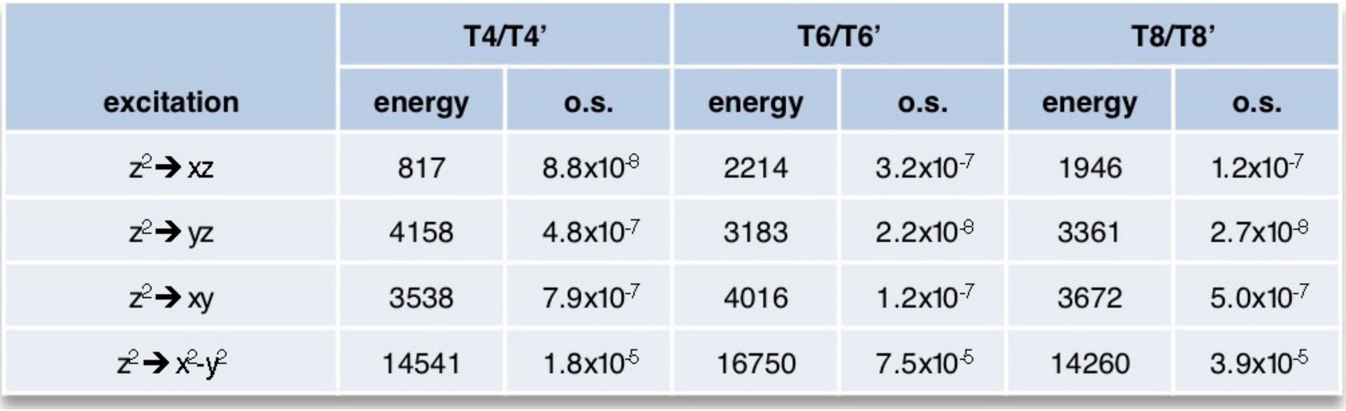


Extended Data Figure 7 | MCD features of CH_4-reacted Fe-BEA. Comparison of MCD data, collected at a temperature of 2.6 K and a field of 7 T, from N_2O-activated Fe-BEA before (black trace) and after (grey trace) reaction with CH_4 at room temperature.



Extended Data Figure 8 | Influence of Fe(III)-OH geometry on O–H bond strength. Shown are models of $S = 5/2$ Fe(III)-OH sites and the associated $S = 2$ Fe(IV)=O species, with O–H bond strengths indicated on the arrows. **a**, Site 1 features a dianionic macrocyclic ligand resembling a β-6MR of a zeolite. **b**, Geometry optimization of the axial oxo structure in **a** shows that this site 1 conformation is destabilized by 4.5 kcal mol^{-1} (or 6.0 kcal mol^{-1}, after correcting for strain of the macrocylic ligand). **c**, Site 2 is bound by two bidentate $[AlH_2(OH)_2]^-$ ligands resembling Al T-sites. **d**, **e**, Sites 3 (**d**) and 4 (**e**) are bound by acac-like bidentate ligands (3-oxo-propenolate).

Extended Data Table 1 | Excitation energies and oscillator strengths for $S=2$ Fe(II) candidate structures



excitation	T4/T4’		T6/T6’		T8/T8’	
	energy	o.s.	energy	o.s.	energy	o.s.
$z^2 \rightarrow xz$	817	$8.8\text{x}10^{-8}$	2214	$3.2\text{x}10^{-7}$	1946	$1.2\text{x}10^{-7}$
$z^2 \rightarrow yz$	4158	$4.8\text{x}10^{-7}$	3183	$2.2\text{x}10^{-8}$	3361	$2.7\text{x}10^{-8}$
$z^2 \rightarrow xy$	3538	$7.9\text{x}10^{-7}$	4016	$1.2\text{x}10^{-7}$	3672	$5.0\text{x}10^{-7}$
$z^2 \rightarrow x^2\text{-}y^2$	14541	$1.8\text{x}10^{-5}$	16750	$7.5\text{x}10^{-5}$	14260	$3.9\text{x}10^{-5}$

The table shows the CASPT2 (refs 8, 11) excitation energies (in cm^{-1}) and corresponding oscillator strengths (o.s.) of ligand-field excited states for the different $S=2$ Fe(II) candidate structures derived from the β-6MR ring A1. In all cases the $3d_{z^2} \rightarrow 3d_{x^2-y^2}$ transition is the highest-energy ligand-field excitation (the *x*- and y-axis are defined by the Fe–O$_{\text{lattice}}$ bonds). This transition has the highest calculated oscillator strength, by one to two orders of magnitude.

LETTER

doi:10.1038/nature19056

Metallaphotoredox-catalysed *sp*^{3}–*sp*^{3} cross-coupling of carboxylic acids with alkyl halides

Craig P. Johnston[1]*, Russell T. Smith[1]*, Simon Allmendinger[1] & David W. C. MacMillan[1]

In the past 50 years, cross-coupling reactions mediated by transition metals have changed the way in which complex organic molecules are synthesized. The predictable and chemoselective nature of these transformations has led to their widespread adoption across many areas of chemical research[1]. However, the construction of a bond between two *sp*^{3}-hybridized carbon atoms, a fundamental unit of organic chemistry, remains an important yet elusive objective for engineering cross-coupling reactions[2]. In comparison to related procedures with *sp*^{2}-hybridized species, the development of methods for *sp*^{3}–*sp*^{3} bond formation via transition metal catalysis has been hampered historically by deleterious side-reactions, such as β-hydride elimination with palladium catalysis or the reluctance of alkyl halides to undergo oxidative addition[3,4]. To address this issue, nickel-catalysed cross-coupling processes can be used to form *sp*^{3}–*sp*^{3} bonds that utilize organometallic nucleophiles and alkyl electrophiles[5–7]. In particular, the coupling of alkyl halides with pre-generated organozinc[8,9], Grignard[10] and organoborane[11] species has been used to furnish diverse molecular structures. However, the manipulations required to produce these activated structures is inefficient, leading to poor step- and atom-economies. Moreover, the operational difficulties associated with making and using these reactive coupling partners, and preserving them through a synthetic sequence, has hindered their widespread adoption. A generically useful *sp*^{3}–*sp*^{3} coupling technology that uses bench-stable, native organic functional groups, without the need for pre-functionalization or substrate derivatization, would therefore be valuable. Here we demonstrate that the synergistic merger of photoredox and nickel catalysis enables the direct formation of *sp*^{3}–*sp*^{3} bonds using only simple carboxylic acids and alkyl halides as the nucleophilic and electrophilic coupling partners, respectively. This metallaphotoredox protocol is suitable for many primary and secondary carboxylic acids. The merit of this coupling strategy is illustrated by the synthesis of the pharmaceutical tirofiban in four steps from commercially available starting materials.

Within the field of drug discovery, there is a demonstrated statistical correlation between clinical success and the molecular complexity of medicinal candidates with respect to the inherent ratio of *sp*^{2}–*sp*^{3} to *sp*^{3}–*sp*^{3} bond content[12]. Not surprisingly, these findings have created an emerging demand within medicinal chemistry for new reaction technologies that enable rapid access to drug-like molecules via the coupling of fragments that incorporate or build novel *sp*^{3}–*sp*^{3} bonds. However, a major hurdle associated with achieving *sp*^{3}–*sp*^{3} bond formation via transition metal catalysis is the limited availability of a diverse suite of nucleophilic coupling partners that are bench-stable, inexpensive, and easily procured. An attractive option would be to use simple carboxylic acids, an abundant native functional group that is chemically robust yet can be readily exploited as a latent leaving group after multistep synthetic sequences (Fig. 1).

The emergence of visible-light-mediated photoredox catalysis within the field of synthetic organic chemistry has enabled the discovery and invention of numerous unique and valuable transformations[13,14]. Indeed, the electronic duality of photocatalyst excited states (which are simultaneously strong oxidants and reductants) has prompted the exploitation of these polypyridyl transition metal complexes in unconventional bond disconnections[15] and facilitated the manipulation of oxidation states in organometallic complexes to enable previously elusive reactions[16–19]. For example, the synergistic merger of single-electron transfer (SET) based decarboxylation with nickel-activated electrophiles has promoted the formation of valuable *sp*^{2}–*sp*^{3} bonds while broadening the field of cross-coupling chemistry via the use of non-conventional reaction substrates[20,21]. This work has further served as a foundation for several extensions of our decarboxylative nickel cross-coupling concept using preactivated phthalimido-derivatized acids[7,22]. Recently, we hypothesized that a straightforward and generic procedure might be developed to enable *sp*^{3}–*sp*^{3} bond formation via the application of ubiquitous carboxylic acids in a decarboxylative cross-coupling with alkyl halides using photoredox catalysis. In developing a method for direct coupling of carboxylic acids with alkyl halides, we hoped to introduce a paradigm for carbon–carbon bond construction that would (i) provide rapid access to complex fragments via *sp*^{3}–*sp*^{3} coupling, (ii) systematically streamline synthetic routes towards drug candidates, and (iii) enable alkyl–alkyl coupling using native functional groups and without substrate preactivation.

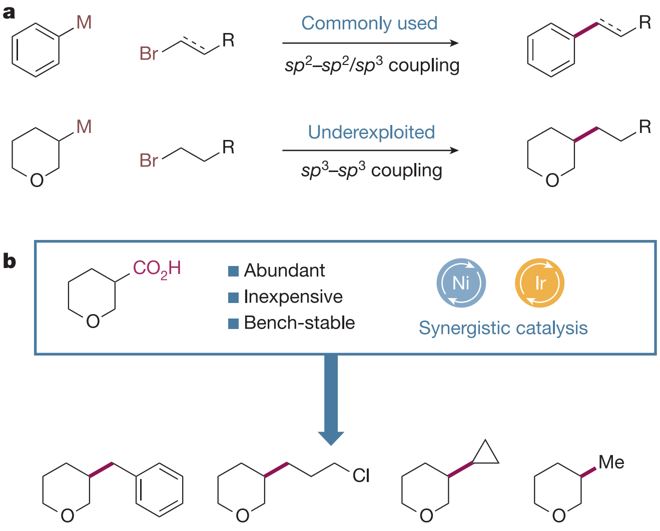


Figure 1 | Carboxylic acids as coupling partners in a metallaphotoredox-mediated process to form *sp*^{3}–*sp*^{3} bonds. a, The majority of transition-metal-catalysed cross-couplings commonly use at least one *sp*^{2}-hybridized coupling partner (top). At present, *sp*^{3}–*sp*^{3} cross-coupling is underexploited (bottom). **b**, The direct utilization of abundant, bench-stable, native functional groups such as carboxylic acids in combination with Ni and Ir synergistic catalysis under mild conditions (top) should encourage greater adoption of *sp*^{3}–*sp*^{3} bond forming methods, leading to products such as those shown at the bottom. M, metal; R, organic functional group.

[1] Merck Center for Catalysis at Princeton University, Princeton, New Jersey 08544, USA.
*These authors contributed equally to this work.

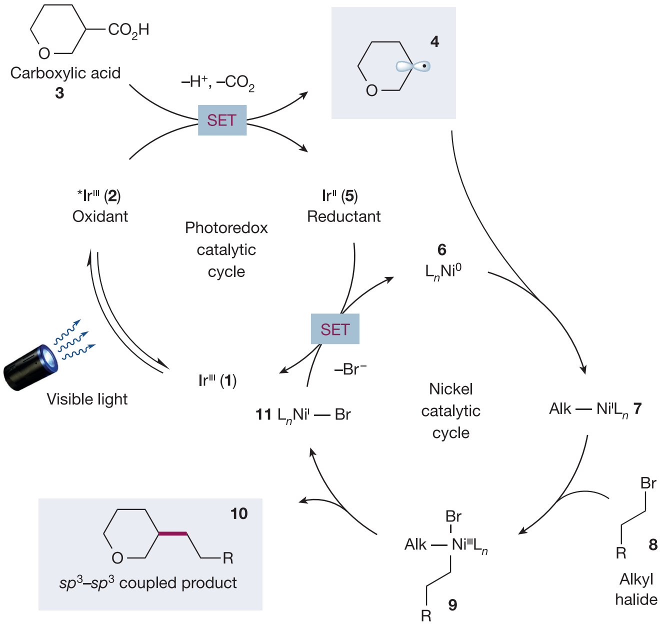


Figure 2 | Proposed mechanism for the metallaphotoredox-mediated cross-coupling of carboxylic acids to generate sp^3–sp^3 bonds. The photoredox catalytic cycle commences with excitation of Ir^{III} **1** to give the excited state **2**. Single electron oxidation of the carboxylate anion derived from acid **3** by oxidant **2** produces the alkyl radical **4** after CO$_2$-extrusion along with Ir^{II} **5**. The nickel catalytic cycle starts with Ni^0 catalyst **6** capturing the alkyl radical **4** (boxed at top) to form the Ni^I species **7**. Ensuing oxidative addition with alkyl halide **8** leads to nickel(III) intermediate **9**. Reductive elimination would then liberate the desired product **10** (boxed at bottom) and Ni^I **11**. Both catalytic cycles converge to complete a single turnover via a SET event that regenerates the photoredox and nickel catalysts.

A detailed mechanism for the proposed decarboxylative sp^3–sp^3 coupling is delineated in Fig. 2. Initial visible-light excitation of the iridium(III) photocatalyst Ir[dF(CF$_3$)ppy]$_2$(dtbbpy)PF$_6$ (**1**) would generate the long-lived (lifetime $\tau = 2.3\,\mu$s)[23] excited-state *Ir^{III} complex **2** (dF(CF$_3$)ppy = 2-(2,4-difluorophenyl)-5-(trifluoromethyl)pyridine, dtbbpy = 4,4′-di-*tert*-butyl- 2,2′-bipyridine). Complex **2** is a strong single-electron oxidant (half-wave redox potential $E_{1/2}^{red}$ [*Ir^{III}/Ir^{II}] = +1.21 V versus the standard calomel electrode, SCE, in CH$_3$CN)[23] and should undergo reduction by a carboxylate anion derived from deprotonation of the acid **3**. The resultant carboxyl radical is expected to rapidly extrude CO$_2$ to produce alkyl radical **4** and the reduced Ir^{II} catalyst **5**. Concurrently, the ligated nickel(0) complex **6** is generated *in situ* via two discrete SET reductions of (dtbbpy)Ni(II)Cl$_2$ by the iridium(II) state of the photocatalyst through sacrificial carboxylic acid consumption ($E_{1/2}^{red}$ [Ir^{III}/Ir^{II}] = −1.37 V versus SCE in CH$_3$CN, $E_{1/2}^{red}$ [Ni^{II}/Ni^0] = − 1.2 V versus SCE in dimethylformamide, DMF)[23,24]. The Ni^0 complex **6** can intercept radical **4** to produce alkyl-Ni^I intermediate **7**[25]. Subsequent oxidative addition with alkyl halide **8** forms the putative organometallic Ni^{III} species **9**, which after reductive elimination forges the desired sp^3–sp^3 bond to furnish the coupled product **10** and Ni^I adduct **11**[26–28]. At this stage the two catalytic cycles would converge by reduction of nickel(I) intermediate **11** by the reduced form of the iridium photocatalyst to re-establish both the Ir^{III} complex **1** and the Ni^0 catalyst **6**[23]. At present, we cannot rule out the possibility of an alternative mechanism that involves Ni^0-mediated oxidative addition and trapping of the alkyl radical **4** by a Ni^{II} species[20,25].

Based on this approach, we began our primary investigations with consideration of the metallaphotoredox conditions previously developed within our group[20]. In these studies we used *N*-Boc proline (Boc, *t*-butyloxycarbonyl) and 1-bromo-3-phenylpropane as coupling partners. Unfortunately, owing to the basic reaction conditions in combination with the polar aprotic solvent DMF, exclusive ester formation was observed. Therefore, we recognized that judicious selection of solvent and base would be necessary to suppress this unwanted by-product formation without obstructing the desired photocatalytic oxidation reaction pathway. To accomplish this goal, a survey of solvents and inorganic bases was conducted in the presence of Ir[dF(CF$_3$)ppy]$_2$(dtbbpy)PF$_6$, NiCl$_2$·glyme and dtbbpy with visible-light irradiation from blue-light-emitting diodes (LEDs). This revealed that the combination of acetonitrile and K$_2$CO$_3$ greatly reduced the rate of carboxylate alkylation and furnished the desired product in 68% yield. Further optimization improved conversion to product through switching to the more electron-rich ligand 4,4′-dOMe-bpy (4,4′-dimethoxy-2,2′-bipyridine). Finally, the addition of water, which decelerated ester formation, provided an additional enhancement and an isolated yield of 85%. A series of control experiments omitting each individual reaction component highlighted the importance of photocatalyst, nickel and light for promoting this decarboxylative sp^3–sp^3 bond-forming reaction (see Supplementary Information).

With optimal metallaphotoredox conditions in hand, we probed the generality of this process with respect to the aliphatic halide electrophile. As representative nucleophilic coupling partners, *N*-Boc- and *N*-Cbz proline (Cbz, carboxybenzyl) were used interchangeably for this purpose (Fig. 3). Simple unfunctionalized primary alkyl halides such as 1-bromo-3-phenylpropane were examined, and these underwent efficient cross-coupling (**12**, 85% yield). Functionality vicinal to the bromide was also tolerated, with benzyl 2-bromoethyl ether coupling in good yield (**13**, 65% yield). A substrate bearing a terminal olefin performed favourably under the reaction conditions to deliver the desired product (**14**, 84% yield). As this protocol is conducted at near ambient temperature, hydrolysis of an ethyl ester does not occur under the optimized basic reaction conditions, and the corresponding carbamate-protected amine was isolated in good yield (**15**, 64% yield). In addition, perfect chemoselective functionalization of an alkyl bromide in the presence of a primary chloroalkane was observed (**16**, 96% yield). A deprotection-cyclization sequence with this material would provide rapid access to the bicyclic tertiary amine core of the naturally occurring pyrrolizidine alkaloids. The presence of free hydroxyl groups is fully compatible with the iridium photocatalyst and the nickel complex (**17**, 86% yield). Moreover, reactive Lewis basic functionalities, such as epoxides and aldehydes, are well tolerated in this cross-coupling procedure and provide numerous opportunities for further derivatization (**18** and **19**, 83% and 62% yield, respectively).

The influence of substitution on the alkyl halide was also investigated to probe the steric limits of the electrophilic coupling partner. No detrimental effects to the efficiency of this process were observed when a β,β-disubstituted bromoalkane was used, and neopentyl bromide coupled in modest yield (**20** and **21**, 75% and 52% yield, respectively). The higher propensity for activated electrophiles to promote esterification led to the utilization of benzyl chloride, as opposed to benzyl bromide, which generated a homobenzylic amine in good yield (**22**, 84% yield). Notably, bromomethane was a competent coupling partner in this protocol which formally affords the product of a fully reduced carboxylic acid moiety in a single step (**23**, 62% yield).

Expanding the substrate scope to encompass secondary alkyl halides permitted us to forge sp^3–sp^3 bonds with adjacent tertiary carbon centres. For example, five- and six-membered cyclic bromoalkanes smoothly reacted to form the desired alkylated products in good to excellent yields (**24**–**26**, 57−91% yield). Smaller ring systems, including cyclopropane and oxetane, were also introduced via this metallaphotoredox procedure (**27** and **28**, 50% and 79% yield, respectively). These motifs have found application in drug discovery programmes as chemically and metabolically stable bioisosteres[29]. Lastly, an acyclic secondary alkyl bromide was also successfully cross-coupled to generate the desired Cbz-protected amine (**29**, 69% yield).

We subsequently examined the scope of the nucleophilic component and found that an assortment of readily available carboxylic acids were viable for this transformation. For instance,

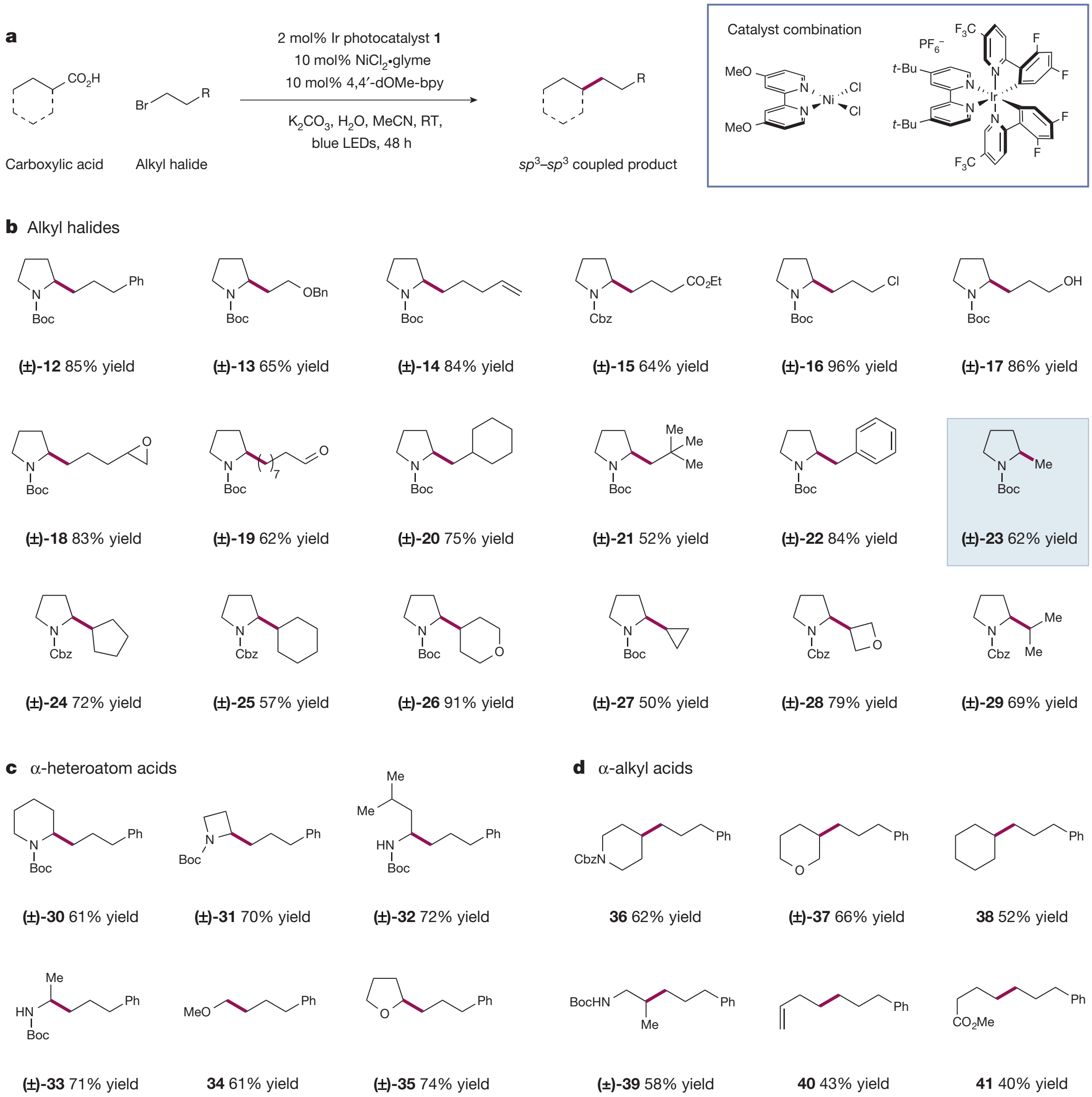


Figure 3 | Carboxylic acid and alkyl halide scope in the dual nickel-photoredox catalysed *sp*^{3}–*sp*^{3} coupling reaction. A broad array of alkyl halides and carboxylic acids are amenable coupling partners in this transformation. **a**, Generalized reaction; **b**–**d**, substrate scope. Optimal catalysts shown in the top right box (some substrates require minor modification; see Supplementary Information). Primary and secondary electrophiles were coupled efficiently with proline derivatives. Alternative α-heteroatom substituted carboxylic acids could also be used to form functionalized amines and ethers. Challenging substrates lacking apparent radical stabilization could also be used successfully. Isolated yields are reported below each entry. See Supplementary Information for full experimental details. For product (±)-**23** shown in shaded box at right in **b**, a yield of 55% was obtained when the reaction was run in flow (GC yield); see Supplementary Information. For product **40** at bottom right in **d**, cyclopropylacetic acid was used.

Boc-protected pipecolic acid and an azetidine derivative both underwent decarboxylative coupling to furnish alkylated products in good yields (**30** and **31**, 61% and 70% yield, respectively). Naturally occurring amino acids, which are inexpensive and obtainable from ample biomass feedstocks, can also be exploited to form α-functionalized amines with excellent efficiency (**32** and **33**, 72% and 71% yield, respectively). Similarly, *O*-methylated glycolic acid functions well to provide access to linear ethers in a straightforward manner (**34**, 61% yield). The cyclic substrate tetrahydrofuran-2-carboxylic acid was also coupled with 1-bromo-3-phenylpropane under these dual nickel-photoredox conditions to afford the ethereal product in very good yield (**35**, 74% yield). Although beneficial, the inclusion of an α-heteroatom on the acid fragment is not a prerequisite for this *sp*^{3}–*sp*^{3} bond forming process. For example, Cbz-protected isonipecotic acid and a tetrahydro-2*H*-pyran derivative were cleanly converted to the corresponding coupled products in an effective fashion (**36** and **37**, 62% and 66% yield, respectively). Simple alkyl precursors lacking heteroatoms can also be used, with cyclohexanecarboxylic acid reacting in reasonable yield (**38**, 52% yield). An acyclic β-amino acid that would generate a secondary radical upon decarboxylation exhibited respectable efficiency and offers the opportunity to synthesize β-functionalized amines with ease (**39**, 58% yield).

Finally, two primary substrates were evaluated in this system to fully exemplify the power of this new *sp*^{3}–*sp*^{3} coupling paradigm. The rearrangement of a cyclopropyl system under the reaction conditions presents the opportunity to produce alkylated homoallylic products in a single step (**40**, 43% yield). Moreover, the monomethyl ester of glutaric acid was subjected to the optimized metallaphotoredox procedure and provided an encouraging quantity of the desired product (**41**, 40% yield).

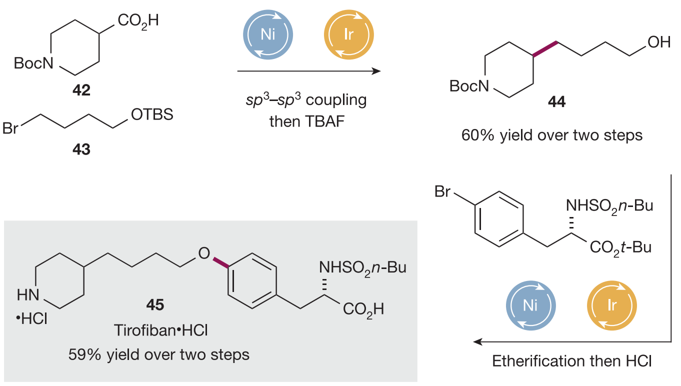


Figure 4 | Application of two metallaphotoredox strategies to the synthesis of tirofiban. The cross-coupling of acid **42** and alkyl halide **43** (top left) using Ni and Ir catalysis generates a new sp^3–sp^3 bond, and subsequent tetrabutylammonium fluoride (TBAF) deprotection provides alcohol **44** (top right). Tirofiban **45** (bottom left) is then synthesized in two further steps in good yield via a Ni/Ir-mediated etherification reaction and acidic deprotection; 34% of the bromide starting material was recovered.

This substrate highlights the potential for downstream modification of latent carboxylates since hydrolysis of the methyl ester would unlock the potential for further sp^3–sp^3 bond formation. Thus, molecules that contain multiple carboxylic acids can function as linch-pin reagents for the rapid assembly of complex molecular architectures. It should be noted that when tertiary acids were applied, the coupled products could be obtained in only limited efficiencies (about 5%–10%). Attempts to improve the yields to synthetically useful levels are continuing.

To further demonstrate the synthetic utility of this decarboxylative coupling protocol, we applied it to the synthesis of the antiplatelet drug tirofiban[30]. As shown in Fig. 4, Boc-isonipecotic acid **42** and alkyl bromide **43** (protected to avoid cyclization to tetrahydrofuran, THF) were exposed to the optimized metallaphotoredox conditions to afford alcohol **44** in good yield, following deprotection of the silyl ether. Thereafter, utilization of the previously established dual photoredox–nickel catalytic etherification reaction enabled direct formation of the desired C–O bond[16]. Following acidic deprotection, tirofiban **45** was synthesized in 59% yield over the final two steps.

We have established a robust strategy for the direct formation of sp^3–sp^3 bonds from abundant carboxylic acids and alkyl halides. This new platform for carbon–carbon bond construction is enabled by the catalytic activation of both coupling partners through the synergistic merger of photoredox and nickel catalysis. The benign nature of the reaction conditions has been exemplified by the breadth of functional groups tolerated in this transformation. We believe that the generality of this methodology and the ready availability of the starting materials used will aid the uptake of sp^3–sp^3 cross-coupling across several fields of synthetic organic chemistry.

Received 21 April; accepted 13 June 2016.

1. de Meijere, A. & Diederich, F. *Metal-catalyzed Cross-coupling Reactions* (Wiley-VCH, 2004).
2. Geist, E., Kirschning, A. & Schmidt, T. sp^3-sp^3 Coupling reactions in the synthesis of natural products and biologically active molecules. *Natural Prod. Rep.* **31,** 441–448 (2014).
3. Haas, D., Hammann, J. M., Greiner, R. & Knochel, P. Recent developments in Negishi cross-coupling reactions. *ACS Catal.* **6,** 1540–1552 (2016).
4. Phapale, V. B. & Cárdenas, D. J. Nickel-catalysed Negishi cross-coupling reactions: scope and mechanisms. *Chem. Soc. Rev.* **38,** 1598–1607 (2009).
5. Jana, R., Pathak, T. P. & Sigman, M. S. Advances in transition metal (Pd,Ni,Fe)-catalyzed cross-coupling reactions using alkyl-organometallics as reaction partners. *Chem. Rev.* **111,** 1417–1492 (2011).
6. Tasker, S. Z., Standley, E. A. & Jamison, T. F. Recent advances in homogeneous nickel catalysis. *Nature* **509,** 299–309 (2014).
7. Qin, T. *et al.* A general alkyl-alkyl cross-coupling enabled by redox-active esters and alkylzinc reagents. *Science* **352,** 801–805 (2016).
8. Giovannini, R., Stüdemann, T., Dussin, G. & Knochel, P. An efficient nickel-catalyzed cross-coupling between sp^3 carbon centers. *Angew. Chem. Int. Ed.* **37,** 2387–2390 (1998).
9. Zhou, J. & Fu, G. C. Cross-couplings of unactivated secondary alkyl halides: room-temperature nickel-catalyzed Negishi reactions of alkyl bromides and iodides. *J. Am. Chem. Soc.* **125,** 14726–14727 (2003).
10. Terao, J., Watanabe, H., Ikumi, A., Kuniyasu, H. & Kambe, N. Nickel-catalyzed cross-coupling reaction of Grignard reagents with alkyl halides and tosylates: remarkable effect of 1,3-butadienes. *J. Am. Chem. Soc.* **124,** 4222–4223 (2002).
11. Saito, B. & Fu, G. C. Alkyl–alkyl Suzuki cross-couplings of unactivated secondary alkyl halides at room temperature. *J. Am. Chem. Soc.* **129,** 9602–9603 (2007).
12. Lovering, F., Bikker, J. & Humblet, C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **52,** 6752–6756 (2009).
13. Narayanam, J. M. R. & Stephenson, C. R. J. Visible light photoredox catalysis: applications in organic synthesis. *Chem. Soc. Rev.* **40,** 102–113 (2011).
14. Prier, C. K., Rankic, D. A. & MacMillan, D. W. C. Visible light photoredox catalysis with transition metal complexes: applications in organic synthesis. *Chem. Rev.* **113,** 5322–5363 (2013).
15. Jeffrey, J. L., Terrett, J. A. & MacMillan, D. W. C. O–H hydrogen bonding promotes H-atom transfer from αC–H bonds for C-alkylation of alcohols. *Science* **349,** 1532–1536 (2015).
16. Terrett, J. A., Cuthbertson, J. D., Shurtleff, V. W. & MacMillan, D. W. C. Switching on elusive organometallic mechanisms with photoredox catalysis. *Nature* **524,** 330–334 (2015).
17. Kalyani, D., McMurtrey, K. B., Neufeldt, S. R. & Sanford, M. S. Room-temperature C–H arylation: merger of Pd-catalyzed C–H functionalization and visible-light photocatalysis. *J. Am. Chem. Soc.* **133,** 18566–18569 (2011).
18. Sahoo, B., Hopkinson, M. N. & Glorius, F. Combining gold and photoredox catalysis: visible light-mediated oxy- and aminoarylation of alkenes. *J. Am. Chem. Soc.* **135,** 5505–5508 (2013).
19. Fabry, D. C., Zoller, J., Raja, S. & Rueping, M. Combining rhodium and photoredox catalysis for C-H functionalizations of arenes: oxidative Heck reactions with visible light. *Angew. Chem. Int. Ed.* **53,** 10228–10231 (2014).
20. Zuo, Z. *et al.* Merging photoredox with nickel catalysis: coupling of α-carboxyl sp^3-carbons with aryl halides. *Science* **345,** 437–440 (2014).
21. Tellis, J. C., Primer, D. N. & Molander, G. A. Single-electron transmetalation in organoboron cross-coupling by photoredox/nickel dual catalysis. *Science* **345,** 433–436 (2014).
22. Huihui, K. M. M. *et al.* Decarboxylative cross-electrophile coupling of *N*-hydroxyphthalimide esters with aryl iodides. *J. Am. Chem. Soc.* **138,** 5016–5019 (2016).
23. Lowry, M. S. *et al.* Single-layer electroluminescent devices and photoinduced hydrogen production from an ionic iridium(III) complex. *Chem. Mater.* **17,** 5712–5719 (2005).
24. Durandetti, M., Devaud, M. & Perichon, J. Investigation of the reductive coupling of aryl halides and/or ethyl chloroacetate electrocatalyzed by the precursor NiX_2(bpy) with $X^- = Cl^-$, Br^- or $MeSO_3^-$ and bpy = 2,2′-dipyridyl. *New J. Chem.* **20,** 659–667 (1996).
25. Gutierrez, O., Tellis, J. C., Primer, D. N., Molander, G. A. & Kozlowski, M. C. Nickel-catalyzed cross-coupling of photoredox-generated radicals: uncovering a general manifold for stereoconvergence in nickel-catalyzed cross-couplings. *J. Am. Chem. Soc.* **137,** 4896–4899 (2015).
26. Lin, X. & Phillips, D. L. Density functional theory studies of Negishi alkyl–alkyl cross-coupling reactions catalyzed by a methylterpyridyl-Ni(I) complex. *J. Org. Chem.* **73,** 3680–3688 (2008).
27. Ren, P., Vechorkin, O., von Allmen, K., Scopelliti, R. & Hu, X. A structure–activity study of Ni-catalyzed alkyl–alkyl Kumada coupling. Improved catalysts for coupling of secondary alkyl halides. *J. Am. Chem. Soc.* **133,** 7084–7095 (2011).
28. Schley, N. D. & Fu, G. C. Nickel-catalyzed Negishi arylations of propargylic bromides: a mechanistic investigation. *J. Am. Chem. Soc.* **136,** 16588–16593 (2014).
29. Meanwell, N. A. Synopsis of some recent tactical application of bioisosteres in drug design. *J. Med. Chem.* **54,** 2529–2591 (2011).
30. Hartman, G. D. *et al.* Non-peptide fibrinogen receptor antagonists. 1. Discovery and design of exosite inhibitors. *J. Med. Chem.* **35,** 4640–4642 (1992).

Supplementary Information is available in the online version of the paper.

Acknowledgements Financial support was provided by NIHGMS (R01 GM093213-01); we acknowledge gifts from Merck, AbbVie, and Bristol-Myers Squibb. C.P.J. thanks Marie Curie Actions for an international outgoing fellowship (PIOF-GA-2013-627695). S.A. thanks the Deutsche Forschungsgemeinschaft (DFG) for a postdoctoral fellowship (AL 1860/2-1).

Author Contributions C.P.J., R.T.S. and S.A. performed and analysed experiments. C.P.J., R.T.S., S.A. and D.W.C.M. designed experiments to develop this reaction and probe its utility, and wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.W.C.M. (dmacmill@princeton.edu).

Reviewer Information *Nature* thanks E. Peterson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

doi:10.1038/nature18594

An early geodynamo driven by exsolution of mantle components from Earth's core

James Badro[1,2], Julien Siebert[1] & Francis Nimmo[3]

Recent palaeomagnetic observations[1] report the existence of a magnetic field on Earth that is at least 3.45 billion years old. Compositional buoyancy caused by inner-core growth[2] is the primary driver of Earth's present-day geodynamo[3–5], but the inner core is too young[6] to explain the existence of a magnetic field before about one billion years ago. Theoretical models[7] propose that the exsolution of magnesium oxide—the major constituent of Earth's mantle—from the core provided a major source of the energy required to drive an early dynamo, but experimental evidence for the incorporation of mantle components into the core has been lacking. Indeed, terrestrial core formation occurred in the early molten Earth by gravitational segregation of immiscible metal and silicate melts, transporting iron-loving (siderophile) elements from the silicate mantle to the metallic core[8–10] and leaving rock-loving (lithophile) mantle components behind. Here we present experiments showing that magnesium oxide dissolves in core-forming iron melt at very high temperatures. Using core-formation models[11], we show that extreme events during Earth's accretion (such as the Moon-forming giant impact[12]) could have contributed large amounts of magnesium to the early core. As the core subsequently cooled, exsolution[7] of buoyant magnesium oxide would have taken place at the core–mantle boundary, generating a substantial amount of gravitational energy as a result of compositional buoyancy. This amount of energy is comparable to, if not more than, that produced by inner-core growth, resolving the conundrum posed by the existence of an ancient magnetic field prior to the formation of the inner core.

At the present day, the geodynamo is powered primarily by compositional buoyancy[3–5] due to the crystallization of the inner core from the outer core, which started around one billion years ago[2,6]. This creates a conundrum as to the origin of the early field; the inner core is certainly much younger than 3.45 Gyr, so a process other than its crystallization must have driven the early field.

Whether an early dynamo could have been driven by thermal buoyancy alone depends on the power extracted from the core by the mantle, which is uncertain[13]. It has been suggested[7,14–16] that light elements dissolved in the core during core formation could have exsolved early in Earth's history as the core cooled; the resulting compositional buoyancy would have generated enough energy to fuel an early geodynamo. Magnesium exsolution before inner-core growth has been proposed[7,15] as a mechanism paralleling oxygen and/or silicon exsolution after inner-core crystallization. The prerequisite however is that magnesium must dissolve in iron during core formation.

To assess the plausibility of that mechanism, we experimentally investigated the solubility of magnesium in molten iron in equilibrium with basaltic and pyrolitic silicate melts at extremely high temperature. The experiments were performed in a laser-heated diamond-anvil cell, in which thin disks of pure iron were sandwiched between two pyrolite or tholeiite glass disks of identical composition, thickness and diameter. The assembly was compressed to 35–74 GPa and laser-heated between 3,300 K and 4,400 K for 30–60 s. After quench and decompression, thin sections were removed from the centre of the laser-heated spot using a crossbeam focused-ion-beam microscope. The thin sections were imaged by high-resolution field-emission scanning electron microscopy and all showed a coalesced spherical iron ball surrounded by molten silicate (Extended Data Fig. 1), confirming that the sample (metal and silicate) was fully molten during equilibration. The composition of the metal and silicate was analysed using high-resolution electron probe microanalysis (see Methods).

Magnesium solubility in iron takes place according to

$$\mathrm{MgO}^{\mathrm{silicate}} \rightleftharpoons \mathrm{Mg}^{\mathrm{metal}} + \mathrm{O}^{\mathrm{metal}} \tag{1}$$

with an equilibrium constant K_{Mg} of $\log(K_{\mathrm{Mg}}) = a + b/T + cP/T$, where T is temperature in kelvin and P is pressure in gigapascals. The parameters (a, b and c; see Methods) were determined from a least-squares fit to our data to obtain

$$\log(K_{\mathrm{Mg}}) = 1.23(0.7) - \frac{18,816(2,600)}{T} \tag{2}$$

where the numbers in parentheses are the standard errors of the parameters. Parameter c was found to be statistically irrelevant (the error on the parameter is larger that the parameter itself and so it does not pass the F test), which indicates that MgO solubility is independent of pressure. The regression is plotted along with the experimental data in Fig. 1 and shows a fit with $R^2 = 0.96$. This confirms that the reaction shown in equation (1) accurately describes the process of MgO dissolution in iron and that pressure has no observable effect. Aluminium dissolution also takes place and can similarly be quantified (Extended Data Fig. 2), as discussed in Methods. However, at extreme temperatures the two-component system reduces to a single homogeneous miscible (solvus) metal-silicate phase[17]. In that case, the reaction shown equation (1) ceases to describe the system because neither of the phases (metal nor silicate) is present. The MgO content of the homogeneous melt is then solely a function of the original bulk composition of the two-phase system.

To estimate the amount of MgO that can be dissolved in the core during formation, we ran a series of multistage core-formation models[11], whereby the planet was grown to its present mass by iterative accretion and core–mantle differentiation of material (see Methods). The magnesium concentrations in the growing core and mantle were calculated iteratively along with other lithophile (O, Si and Al) and siderophile (Ni, Co, Cr and V) elements. More than 8,000 simulations were performed to sample the parameter space fully, and only geochemically consistent models (for which the final concentrations of Ni, Co, Cr and V in the silicate match present-day mantle abundances) were retained[11].

For core formation without a giant impact, we found a maximum of 0.8 wt% MgO in the core, in the most favourable (hottest geotherm and deepest magma ocean) case. For a present-day temperature[18] at the core–mantle boundary (CMB) of 4,100 K, the MgO equilibrium

[1]Institut de Physique du Globe de Paris, Université Sorbonne Paris Cité, 75005 Paris, France. [2]Earth and Planetary Science Laboratory, École Polytechnique Fédérale de Lausanne, CH-1015, Lausanne, Switzerland. [3]Department of Earth and Planetary Sciences, University of California Santa Cruz, Santa Cruz, California 95064, USA.

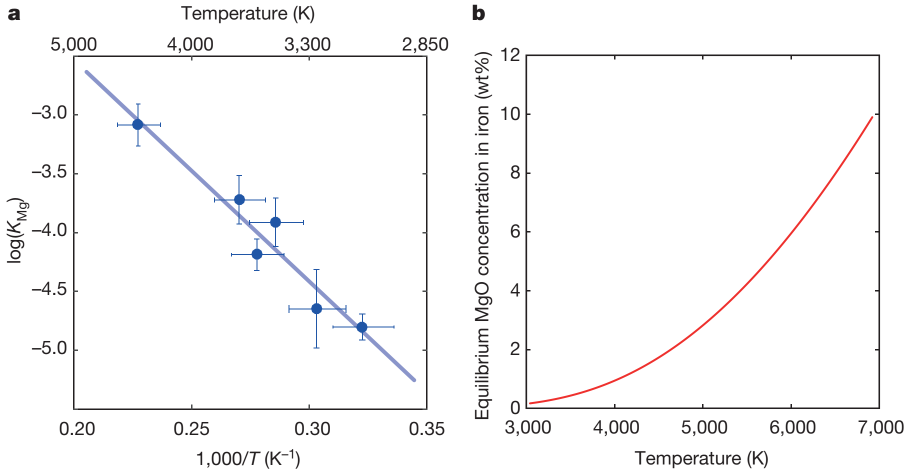


Figure 1 | Magnesium solubility in metallic iron melt at high pressure and temperature. **a**, Equilibrium constant for MgO dissolution in iron (K_{Mg}) as a function of reciprocal temperature (1,000/T) (equation (2)). The experimental data are from Extended Data Table 1. The line corresponds to the least-squares linear fit to the data, and the error bars correspond to 1σ uncertainties. A comparison with extrapolation from density functional theory (DFT) calculations[17] is shown in Extended Data Fig. 2. **b**, The resulting MgO concentration in iron in equilibrium with pyrolite as a function of temperature. This is obtained by rewriting equation (2) to obtain $X_{Mg}^{metal} = 2.91\exp(-21,662/T)$, where T is temperature in kelvin, and then converting Mg molar fractions to MgO weight fractions. For an extended version of this graph, see Extended Data Fig. 6.

value (saturation threshold) in the core is 1.1 wt% (Fig. 1b). The core is therefore under-saturated in MgO so that any primordial magnesium dissolved during formation would not exsolve to the mantle.

We then ran a series of core-formation models that involve a final giant impact. In the Moon-forming giant-impact scenario[12], the impactor is typically thought of as a Mars-sized planetary embryo, but the masses used in models range from 2.5% to 20% of Earth's mass[19,20]. With such a size, the impactor is a differentiated object with a core and mantle, and the temperatures during the impact are sufficiently high that the impactor core and the surrounding silicate mantle turn into a single miscible metal-silicate phase (see Methods). As this dense silicate-saturated metallic object (hereafter called the 'hybridized impactor core', HIC) merges with Earth's core, it strongly increases the lithophile-element content of Earth's core. We calculated the composition of the HIC as a function of impactor size (Fig. 2a) by assessing its dilution ratio[21] (see Methods) in the magma ocean, that is, the relative mass of magma ocean with which the impactor core interacts. The amounts of Mg, Si and O brought by the HIC to Earth's core are plotted in Fig. 2b.

The total MgO dissolved in the core (Extended Data Fig. 3) ranges between 1.6 wt% and 3.6 wt%. Those values are higher than the saturation value of 1.1 wt% at the present-day CMB, which implies that the core became over-saturated in MgO as it cooled. The excess MgO must have exsolved to the mantle and provided a large source of potential energy[7] to drive an early dynamo. Because MgO solubility depends on temperature, but not on pressure, MgO exsolution in the core takes place at the CMB, where the temperature is lowest. As MgO exsolves from the metal, the residue becomes denser and sinks, and is replaced by lighter MgO-bearing metal. This process ensures that the entire core is processed at the CMB, so that the equilibrium concentration at the CMB (Fig. 1b) sets the concentration in the whole core. We estimated the energy released by MgO exsolution by calculating the difference in gravitational energies (ΔE_{grav}) of the core before and after exsolution, with the gravitational energy in each state given by

$$E_{grav} = -\int_0^R \frac{GM(r)}{r} 4\pi r^2 \rho(r) dr \qquad (3)$$

in which G is the gravitational constant, $M(r)$ is the mass of the core within a radius r, $\rho(r)$ is the density of the core at radius r and R is the radius of the core.

The energy release depends on how the HIC mixes with Earth's core, as shown by the dependence on $\rho(r)$ in equation (3). We investigated two extreme models of mixing: (i) full mixing of the HIC with Earth's core producing a homogeneous core and (ii) full layering whereby the HIC sits atop Earth's core (see Methods). The energy release as a function of impactor size is plotted in Fig. 3. In the mixed case, the energy release yields (1–5.5) × 10^{29} J. For comparison, the total energy release from inner-core growth (latent heat and buoyancy) is[2,6] (0.9–1.7) × 10^{29} J. The layered model provides less energy for small impacts (Fig. 3), but again reaches and exceeds the energy released by inner-core growth for Mars-sized or larger impactors.

Because MgO solubility depends only on temperature, the power release and onset time of MgO exsolution depend on the temperature evolution at the CMB, which itself depends on the initial MgO concentration in the core (see Methods). Although the early evolution of CMB temperature is uncertain, as an example we adopt an *a priori* CMB temperature[18] model. Prior to inner-core growth, the exsolution rate is high, as shown in Extended Data Fig. 4, and generates power in excess of about 3 TW (a conservative estimate of the amount of power

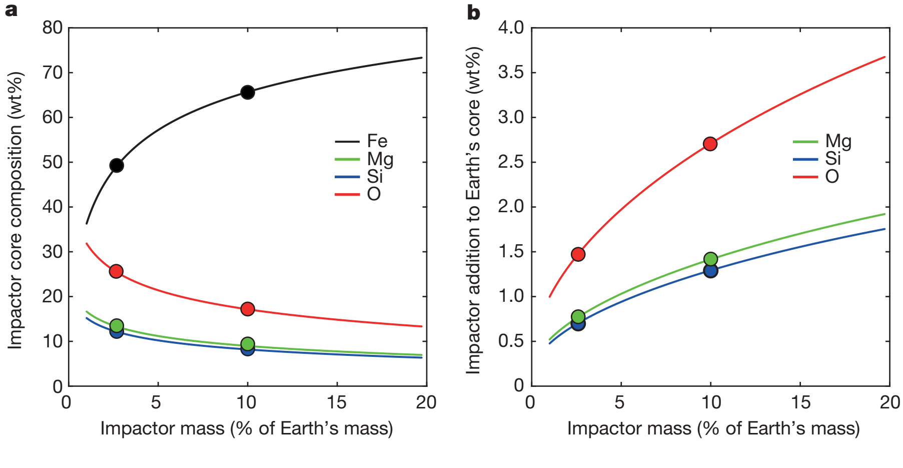


Figure 2 | Composition of the core of the giant impactor after equilibration in the magma ocean, and its effect on Earth's core composition. **a**, The composition of the hybridized impactor core (HIC) plotted as a function of impactor mass. Smaller impactors interact and equilibrate with larger relative amounts of magma ocean material; they 'swell' (see Extended Data Fig. 7) and become very enriched in Mg, Si and O. **b**, The compositional imprint of the giant impact on the core; between 2% and 8% of the total mass of the core consists of mantle material transported by the HIC. The Si and O concentrations added to the core are lower than the amounts present in the core before the impact[23]. This shows that the major contribution of the giant impact to core chemistry is the magnesium influx. The Mars-size impact[19] (10% of Earth's mass) and 'fast-spinning' impact[20] (2.5% of Earth's mass) are highlighted by circles in both panels.

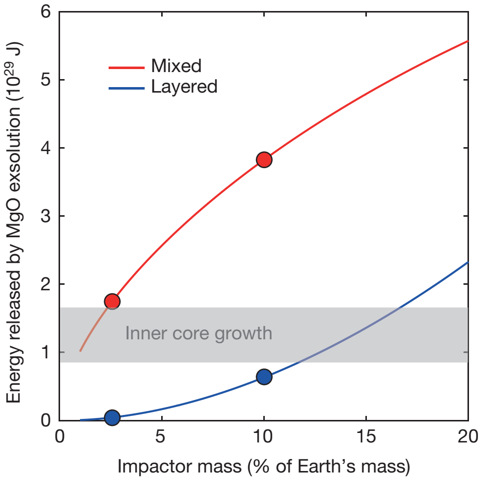


Figure 3 | Gravitational energy released by the exsolution of buoyant mantle components from the core after the giant impact. Calculated following equation (3), the red curve corresponds to the energy released if the HIC fully mixes with Earth's core. In that case, MgO exsolution occurs up to the current saturation limit (Fig. 1b). The blue curve corresponds to the energy released if the HIC forms a layer on top of Earth's core. In that case, the layer is so rich in lithophile elements (Fig. 2a) that the exsolution of all dissolved mantle components (MgO and SiO_2) takes place. The Mars-size impact[19] (10% of Earth's mass) and fast-spinning impact[20] (2.5% of Earth's mass) are highlighted by circles. The grey horizontal band corresponds to the energy released by inner-core growth (gravitational + latent heat) since its inception, and is the main driver for the geodynamo today. The energies released by MgO exsolution are of the order of, if not higher than, those released by inner-core growth, and demonstrate the effectiveness of the exsolution of mantle components to drive an early dynamo. The average power of exsolution can be estimated assuming an exsolution time (Extended Data Fig. 8) or a temperature evolution model of the core (Extended Data Fig. 5).

required to run a geodynamo by compositional buoyancy[22]) over the course of exsolution (Extended Data Fig. 5). With the onset of inner-core growth, the cooling and exsolution rates decrease and the power drops to about 1 TW (see Methods). In terms of timing, the onset of exsolution occurs once the (decreasing) MgO saturation value at the CMB reaches the concentration in the core (Extended Data Fig. 5). For our nominal model, this occurs around 1 Gyr after Earth's formation with a Mars-sized impact and increases to approximately 2.3 Gyr in the case of a small, 'fast-spinning' impact (see Methods).

Rapid initial cooling following a giant impact may have driven an early thermal dynamo. However, our experimental results show that MgO exsolution probably dominated the energy budget of Earth's core in the intermediate period between early, rapid cooling and the onset of inner-core growth. This result provides a tangible basis for an exsolution-driven dynamo[7], as well as a plausible mechanism for explaining the uninterrupted geological record of magnetism[1] in Earth's rocks and minerals dating to 3.5 Gyr ago or earlier. This mechanism should be relatively ineffective in smaller planets such as Mars or on Earth-sized planets that have not experienced a giant impact; but for super-Earths, where pressures and temperatures could remain super-solvus for extended periods, it represents a new method of driving potentially detectable present-day dynamos.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 5 January; accepted 16 May 2016.
Published online 18 July 2016.

1. Tarduno, J. A. *et al.* Geodynamo, solar wind, and magnetopause 3.4 to 3.45 billion years ago. *Science* **327,** 1238–1240 (2010).
2. Nimmo, F. in *Treatise on Geophysics* 2nd edn, Vol. 8 (ed. Olson, P.) 27–55 (Elsevier, 2015).
3. Lister, J. R. & Buffett, B. A. The strength and efficiency of thermal and compositional convection in the geodynamo. *Phys. Earth Planet. Inter.* **91,** 17–30 (1995).
4. Labrosse, S., Poirier, J. P. & LeMouel, J. L. On cooling of the Earth's core. *Phys. Earth Planet. Inter.* **99,** 1–17 (1997).
5. Gubbins, D., Alfe, D., Masters, G., Price, G. D. & Gillan, M. J. Can the Earth's dynamo run on heat alone? *Geophys. J. Int.* **155,** 609–622 (2003).
6. Labrosse, S. Thermal evolution of the core with a high thermal conductivity. *Phys. Earth Planet. Inter.* **247,** 36–55 (2015).
7. O'Rourke, J. G. & Stevenson, D. J. Powering Earth's dynamo with magnesium precipitation from the core. *Nature* **529,** 387–389 (2016).
8. Ringwood, A. E. Chemical evolution of terrestrial planets. *Geochim. Cosmochim. Acta* **30,** 41–104 (1966).
9. Rubie, D. C., Melosh, H. J., Reid, J. E., Liebske, C. & Righter, K. Mechanisms of metal-silicate equilibration in the terrestrial magma ocean. *Earth Planet. Sci. Lett.* **205,** 239–255 (2003).
10. Wood, B. J., Walter, M. J. & Wade, J. Accretion of the Earth and segregation of its core. *Nature* **441,** 825–833 (2006).
11. Badro, J., Brodholt, J. P., Piet, H., Siebert, J. & Ryerson, F. J. Core formation and core composition from coupled geochemical and geophysical constraints. *Proc. Natl Acad. Sci. USA* **112,** 12310–12314 (2015).
12. Hartmann, W. K. & Davis, D. R. Satellite-sized planetesimals and lunar origin. *Icarus* **24,** 504–515 (1975).
13. Labrosse, S., Hernlund, J. W. & Coltice, N. A crystallizing dense magma ocean at the base of the Earth's mantle. *Nature* **450,** 866–869 (2007).
14. Buffett, B. A., Garnero, E. J. & Jeanloz, R. Sediments at the top of Earth's core. *Science* **290,** 1338–1342 (2000).
15. Stevenson, D. Core exsolution: a likely consequence of giant impacts and a likely energy source for the geodynamo. *Eos Trans. AGU* **88** (Fall Meet. Suppl.), abstr. U21D-02 (American Geophysical Union, 2007).
16. Buffett, B. A. Earth's core and the geodynamo. *Science* **288,** 2007–2012 (2000).
17. Wahl, S. M. & Militzer, B. High-temperature miscibility of iron and rock during terrestrial planet formation. *Earth Planet. Sci. Lett.* **410,** 25–33 (2015).
18. Nimmo, F. in *Treatise on Geophysics* 2nd edn, Vol. 9 (ed. Stevenson, D. J.) 201–219 (Elsevier, 2015).
19. Canup, R. M. Forming a Moon with an Earth-like composition via a giant impact. *Science* **338,** 1052–1055 (2012).
20. Cuk, M. & Stewart, S. T. Making the Moon from a fast-spinning earth: a giant impact followed by resonant despinning. *Science* **338,** 1047–1052 (2012).
21. Deguen, R., Landeau, M. & Olson, P. Turbulent metal-silicate mixing, fragmentation, and equilibration in magma oceans. *Earth Planet. Sci. Lett.* **391,** 274–287 (2014).
22. Aubert, J., Labrosse, S. & Poitou, C. Modelling the palaeo-evolution of the geodynamo. *Geophys. J. Int.* **179,** 1414–1428 (2009).
23. Badro, J., Cote, A. S. & Brodholt, J. P. A seismologically consistent compositional model of Earth's core. *Proc. Natl Acad. Sci. USA* **111,** 7542–7545 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 207467. Parts of this work were supported by the UnivEarthS Labex programme at Sorbonne Paris Cité (ANR-10-LABX-0023 and ANR-11-IDEX-0005-02), IPGP multidisciplinary programme PARI, and by Paris–IdF region SESAME grant no. 12015908. J.S. acknowledges support from the French National Research Agency (ANR project VolTerre, grant no. ANR-14-CE33-0017-01). We thank J. Aubert, S. Stewart and P. Asimow for discussions. We thank R. Ryerson for comments on the manuscript.

Author Contributions J.B. designed the project, performed the experiments, implemented the thermodynamic and core-formation modelling, discussed the results and wrote the manuscript. J.S. performed the experiments, discussed the results and commented on the manuscript. F.N. implemented the core exsolution energy modelling, discussed the results and commented on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.B. (badro@ipgp.fr).

Reviewer Information *Nature* thanks B. Buffett and Y. Fei for their contribution to the peer review of this work.

METHODS

Magnesium and aluminium solubility. The thermodynamic process of lithophile element incorporation in iron involves the solubility of mantle components in the metal phase (equation (1)), rather than redox exchange as in the case of siderophile element partitioning. The magnesium concentration in the metal ranges between 0.2 mol% and 1 mol% in our experiments. The equilibrium constant of the dissolution reaction given in equation (1) (reprinted here for convenience):

$$\mathrm{MgO}^{\mathrm{silicate}} \rightleftharpoons \mathrm{Mg}^{\mathrm{metal}} + \mathrm{O}^{\mathrm{metal}}$$

is

$$K_{\mathrm{Mg}} = \frac{X_{\mathrm{Mg}}^{2}}{X_{\mathrm{MgO}}} \qquad (4)$$

(where X_{Mg} is the mole fraction of Mg in the metal and X_{MgO} the mole fraction of MgO in the silicate) and its logarithm is proportional to the change in Gibbs free energy of the reaction defined by equation (1):

$$\log(K_{\mathrm{Mg}}) = a + \frac{b}{T} + c\frac{P}{T} \qquad (5)$$

where the parameters *a*, *b* and *c* correspond to the changes in entropy, enthalpy and volume in the reaction in equation (1), respectively. These parameters were fitted to the data using linear regression, and *c* was found to be statistically irrelevant (no pressure dependence), yielding equation (2) (reprinted here for convenience):

$$\log(K_{\mathrm{Mg}}) = 1.23(\ 0.7) - \frac{18,816(2,600)}{T}$$

Similarly, the aluminium concentration of in the metal ranges from 0 mol% (below the detection limit, explaining two fewer points for the Al plot in Extended Data Fig. 2) to 1.1 mol%. The equilibrium constant of the dissolution reaction

$$\mathrm{Al_2O_3}^{\mathrm{silicate}} \rightleftharpoons 2\mathrm{Al}^{\mathrm{metal}} + 3\mathrm{O}^{\mathrm{metal}} \qquad (6)$$

is

$$K_{\mathrm{Al}} = \frac{X_{\mathrm{Al}}^{2.5}}{X_{\mathrm{AlO_{1.5}}}} \qquad (7)$$

(where X_{Al} is the mole fractions of Al in the metal and $X_{\mathrm{AlO_{1.5}}}$ the mole fraction of $\mathrm{AlO_{1.5}}$ in the silicate) Its logarithm is proportional to the change in Gibbs free energy of the reaction in equation (6) and can be written in the same form as equation (5); fitting to the data using linear regression shows that *c* is once again statistically irrelevant, and we find

$$\log(K_{\mathrm{Al}}) = 4.1(1.4) - \frac{36,469(5,260)}{T} \qquad (8)$$

where the numbers in parentheses are the standard errors of the parameters.

Saturation conditions at the core–mantle boundary. Equations (2), (4), (7) and (8) allow us to calculate the Mg and Al concentrations in molten iron as a function of temperature and silicate composition. An important case is that of the equilibrium value in the core at the core–mantle boundary (CMB). As shown above, MgO dissolution in iron has no pressure dependence. This means that MgO exsolves in the coldest part of the core, which is the CMB. The equilibrium value at the CMB is therefore the MgO saturation value; if the MgO concentration in the core is above saturation, then MgO will be exsolved until it reaches that value. Figure 1b shows the equilibrium value of MgO concentration in the core as a function of CMB temperature, for a core buffered by (that is, in local equilibrium with) a pyrolitic magma ocean (50 mol% MgO in the mantle).

Experimental and analytical. The silicate glasses were produced in an aerodynamic levitation laser furnace. The starting mixes were made by grinding and mixing from pure oxide ($\mathrm{SiO_2}$, MgO, $\mathrm{Fe_2O_3}$, $\mathrm{Al_2O_3}$) and carbonate ($\mathrm{CaCO_3}$) components, pressing them into pellets, and then fusing them at constant oxygen fugacity at 1,900–2,100 °C for 5 min in a laser furnace using a 120-W $\mathrm{CO_2}$ laser. The fused samples were quenched to glasses, and analysed for recrystallization, homogeneity and composition on a Zeiss Auriga field-emission scanning electron microscope (IPGP, Paris). The glass beads were thinned down to 20-μm-thick double-parallel thin sections and were processed using a femtosecond laser machining platform to cut disks of identical size for loading in the diamond-anvil cell. Spherical iron balls 1–3 μm in size were flattened between two such silicate disks, and constituted the layered starting sample. Pressure was measured from the frequency shift of the first-order Raman mode in diamond, measured on the anvil tips. Temperature was measured every second, simultaneously from both sides, by spectroradiometry. Electronic laser shutdown operates in about 2–4 μs, and temperature quench occurs in approximately 10 μs (owing to thermal diffusion in the sample) ensuring an ultrafast quench of the sample.

After decompression, a thin section (20 μm × 10 μm wide, 1–3-μm thick) was extracted from the centre of the laser-heated spot using a Zeiss Auriga crossbeam focused-ion-beam microscope (IPGP, Paris). The sample was imaged and then transferred to a TEM copper grid, and the metal and silicate phases were analysed using a Cameca SX-Five electron microprobe (CAMPARIS, Paris) with five large-area analysers. Metal and silicate phases of the run products are large enough (>5 μm) to perform reliable analyses with an electron probe micro-analyser (EPMA) on focused ion beam (FIB) thin sections.

Metal and silicate phases were analysed using Cameca SX100 and Cameca SX FIVE (CamParis, UPMC–IPGP) electron probe micro-analysers. X-ray intensities were reduced using the CITZAF correction routine. Operating conditions were 15-kV accelerating voltage, and 10–20-nA beam current and counting times of 10–20 s on peak and background for major elements and 20–40 s for trace elements (including Mg and Al in the metallic phases). Pure Fe metal was used as standard for metal. $\mathrm{Fe_2O_3}$, $\mathrm{SiO_2}$, MgO and $\mathrm{Al_2O_3}$ were used as standards to measure solubility of oxygen, silicon, magnesium and aluminium in metal. Diopside glass (Si), wollastonite (Ca), orthoclase (K), anorthite (Al), albite (Na), rutile (Ti) and pure oxides ($\mathrm{Fe_2O_3}$, MgO, $\mathrm{SiO_2}$, CaO and $\mathrm{Al_2O_3}$) were used as standards for the silicate. We verified that the geometry of the metal and silicate phases was identical from both sides of the FIB sections, so that the EPMA analyses only a single phase. The EPMA uses a beam size of 1–2 μm, which is large enough to integrate the small quench features of metal and silicate phases (<200 nm) and to determine their bulk compositions. When a few small metallic blobs were present in the silicate (500 nm to 2 μm in diameter), special care was taken to avoid them during analysis of the silicates.

Core formation modelling. The core of Earth formed in the first approximately 50 million years[24,25] of the Solar System, by an iterative addition of material to the proto-Earth. The accreting material, consisting of mixtures of iron-rich metals and silicates similar to those found in extra-terrestrial bodies (such as chondrite parent bodies, HEDs and angrites), impacted the growing planet. The heat generated by those impacts maintained the outermost portion of the planet in a molten state known as a magma ocean[26]. At temperatures below the solvus of iron and silicate, the two phases un-mix and the metal (twice denser) segregates towards the centre and forms the core. Along with the segregating metal, the siderophile elements are stripped to the core, among which are light elements such as Si and O. The depletion of siderophile elements from the mantle has been widely used to constrain the pressure–temperature–composition path of core formation, and has shown that the core formed in a deep magma ocean[27,28]. As the planet accretes, the magma ocean grows deeper; recent models[11] show that the concentrations of Ni, Co, Cr and V in the mantle satisfy terrestrial observables for a final magma ocean depth of between 1,000 km and 1,700 km, corresponding to final pressures of between 40 GPa and 75 GPa and final temperatures of between 3,000 K and 4,180 K, respectively.

We ran a series of traditional multistage core-formation models[11] where the planet was accreted to its present mass in increments of 0.1% of Earth's mass, without giant impacts. At each stage, the planet grows and the pressure and temperature of equilibration increase accordingly. The concentrations of Ni, Co, V, Cr, O, Si and Mg in the core were calculated iteratively during the 1,000 steps of the accretion process. The simulations were run for a variety of redox paths (ranging from very reduced to very oxidized), several geotherms (between the solidus and the liquidus of peridotite), and for all possible magma ocean depths, ranging from 0% (magma lake) to 100% (fully molten Earth) of the mantle. We forward-propagated all uncertainties on the thermodynamic parameters governing the partitioning equations using Monte Carlo simulation. Most models (very deep or very shallow) do not satisfy, within uncertainties, the observed geochemical abundances of Ni, Co, V and Cr in the mantle and therefore are not relevant. We selected only the models that do reproduce the geochemical abundances of Ni, Co, V and Cr in the present-day mantle, and found that the maximum MgO concentration in the core at the end of accretion is 0.8 wt%.

Giant-impact modelling. In the Moon-forming giant-impact scenario[12], the impactor is typically thought of as a Mars-sized planetary embryo, but the masses used in models range from 2.5% to 20% of Earth's mass[19,20]. With such a size, the impactor is a differentiated object with a core and mantle (as opposed to small undifferentiated bodies) and, hence, it does not fully equilibrate with the entire magma ocean, but rather partially equilibrates[29] with a small portion[9,21] of that magma ocean. The impactor and the magma ocean (in the impact zone) reach tremendous temperatures during the impact, as shown by smoothed-particle hydrodynamic simulations[19,20]. Even though the temperatures from those simulations can be inaccurate because of intrinsic inaccuracies in the equations of state that they are based on, the minimum temperature[19] for the impactor core is

8,000 K and that of the magma ocean in the impacted area is 7,000 K. Therefore, the system consisting of the impactor core and the surrounding silicate mantle is necessarily always hotter than 7,000 K, and turns into a single miscible metal-silicate phase.

We calculated the composition of Earth's core after the giant impact in two steps. First, we modelled the pre-giant-impact accretionary phase. The Earth was partially accreted, as described in the previous paragraph, until it reached 80% to 99% of Earth's mass, leaving the planet in the state it was in before the giant impact. We considered only the models that reproduce the present-day geochemical abundances of Ni, Co, V and Cr in the mantle. Then the final accretion event took place, consisting of the giant impact bringing in the remaining 1% to 20% of Earth's mass. We calculated the composition of the hybridized impactor core (HIC) as a function of its size (Fig. 2a) by considering the fact that, as opposed to small accretionary building blocks, the core of the giant impactor does not fully equilibrate with the entire magma ocean; instead, it partially equilibrates[29] with a small portion[9,21] of the magma ocean (see Methods section 'Partial core equilibration and turbulent fragmentation and mixing' below). It is clear from Fig. 2a that the bigger the impactor, the smaller the relative mass of magma ocean it interacts and equilibrates with and, consequently, the less mantle components (Mg, O, Si) the HIC contains. The net effect on Earth's core, once the HIC is added, is mitigated as shown in Fig. 2b; it is the result of the balance between larger HICs being less enriched in mantle component, but contributing more mass to the whole core.

Partial core equilibration and turbulent fragmentation and mixing. The composition of the HIC was calculated by taking into account two main parameters that are usually neglected in traditional core-formation models[9–11,28,30].

First, the degree of partial equilibration—that is, the fraction of the core that equilibrates with the mantle—has been constrained by geochemical modelling, from the combined analysis of the Hf–W and U–Pb isotopic systems, and shown to be at least[25,29,31] 40%. We used this conservative lower bound, meaning that 60% of the impactor core merges with Earth's core without equilibration (and therefore with no compositional effect), whereas the other half equilibrates in the magma ocean before merging with the core.

Second, the impactor core only 'sees' a portion[9] of the magma ocean, with the fraction involved in the equilibration estimated from fragmentation and turbulent mixing scaling laws[21]; these laws show that the ratio of equilibrated silicate to equilibrated metal (dilution ratio Δ) in the magma ocean is given by

$$\Delta = \frac{\rho_{\text{silicate}}}{\rho_{\text{metal}}}\left[\left(1+\frac{0.25}{\delta^{1/3}}\right)^3 - 1\right]$$

where ρ_{silicate} and ρ_{metal} are the densities of silicate and metal, and δ is the ratio of impactor to Earth mass.

MgO exsolution energy. The energy release depends on how the HIC mixes with Earth's core, as shown by the dependence on $\rho(r)$ in equation (3). Even though simulations[20] and energetic arguments[32] suggest that the HIC should thoroughly mix with Earth's core, we investigated two extreme models of mixing: (i) full mixing of the HIC with Earth's core producing a homogeneous core and (ii) full layering where the HIC sits atop Earth's core.

In the mixed case, the HIC is diluted in the bulk of Earth's core and therefore the Si and O content delivered by the impactor are below the saturation limit of those elements[11,30,33] (Fig. 2b); those concentrations are under-saturated with respect to the overlying conditions imposed by the magma ocean at the CMB, and there is no chemical drive to force those components out of the system. In that case, we considered that MgO is the only phase to exsolve so that the associated energy release is a conservative lower bound.

In the layered case, the HIC is concentrated atop the proto-core, and all three mantle components (MgO, SiO_2 and FeO) are highly concentrated in the layer and over-saturated with respect to CMB conditions prevailing atop that layer. In that case, all of those components would exsolve and remix with the overlying magma ocean.

In our energy calculations, we fixed the present-day CMB temperature to 4,100 K. Lower temperatures imply a lower saturation level in the core, and mean that more MgO exsolves and more energy is produced, and vice versa. The final density and radius of the core are the present-day values (10.6 g cm^{-3} and 3,485 km, respectively).

Impactor core mixing. We considered a uniform core of density ρ and radius R; it subsequently undergoes un-mixing into an inner (dense) region with density ρ_c and radius R_c (the present-day values given above), and an outer buoyant layer with density ρ_{layer}. The volume fraction of the outer layer is f, which we take to be $\ll$1. We may write

$$\rho = (1-f)\rho_c + f\rho_{layer} \tag{9}$$

and

$$R_c = (1-f/3)R \tag{10}$$

where equation (10) is correct to first order in f. In practice, we specify ρ_c and ρ_{layer} (4.8 g cm^{-3} for MgO) and calculate ρ and R for a given value of f, with the current core boundary taken to be R_c. The gravitational energy E of the core in either state may be derived using equation (3), and the change in energy ΔE in going from the uniform to the unmixed state can be available to do work (for example, to drive a dynamo). Making use of equations (3), (9) and (10), it may be shown that, to first order in f:

$$\Delta E = \frac{16}{45}\pi^2 G R^5 f \rho_c(\rho_c - \rho_{\text{layer}}) \tag{11}$$

For $f = 20\%$, equation (11) overestimates the full calculation (plotted in the figures) by about 5%; the discrepancy is smaller with smaller f, and equation (11) can be used, to a good approximation, to estimate the amount of energy released by mantle-component exsolution from the core. This equation shows the correct limiting behaviour in the cases of $f = 0$ and $\rho_c = \rho_{\text{layer}}$.

Impactor core layering. In this case we take the mass fraction of the Earth's core added by the HIC to be f_{m}. With a HIC density of ρ_{i} and a present-day total core mass of M_{c}, the radius of the base of the impactor layer R_1 before un-mixing of this layer is given by

$$R_1^3 = R^3 - \frac{3f_{\text{m}}M_{\text{c}}}{4\pi\rho_{\text{i}}}$$

The HIC layer then undergoes un-mixing into two components: 'mantle components' (ρ_2, 5.6 g cm^{-3}) and 'core material' (ρ_1, 10.6 g cm^{-3}). The HIC density ρ_{i} may then be derived using

$$(R^3 - R_1^3)\rho_{\text{i}} = (R_2^3 - R_1^3)\rho_1 + (R^3 - R_2^3)\rho_2$$

where R_2 is the radius of the base of the light-element layer after un-mixing. To make the total core mass correct, the density of the pre-impact core, ρ_{c}, is also calculated. Once ρ_{i}, R_1 and R_2 have been calculated, the energy change due to un-mixing within the layer can be calculated using successive applications of equation (11), as before.

Thermal evolution and exsolution power. Using a CMB temperature evolution model, we can estimate the MgO exsolution rate and, hence, an exsolution power, as a function of time. A typical CMB temperature evolution is shown in Extended Data Fig. 4a, along with the associated MgO content of the core (Extended Data Fig. 4b) obtained by rewriting the MgO equilibrium curve (Fig. 1b) as a function of time. The time derivatives are the cooling rate of the core and its MgO exsolution rate as a function of time, and are plotted in Extended Data Fig. 4c, d, respectively.

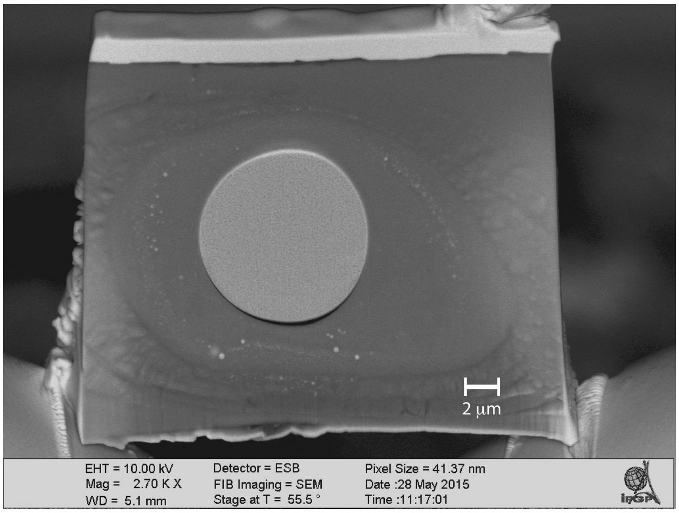
Very early in Earth's history, the core was so hot that the equilibrium MgO concentration at the CMB (Extended Data Fig. 4b) is higher than the MgO content of the core, and no exsolution occurs. The reverse reaction—that is, the potential for MgO to be dissolved from the mantle to the core—is limited; it is prone to affect only a thin layer below the CMB that is enriched in MgO, that becomes light and stably stratified, and that is therefore unable to recycle and affect the entire core. As the core cools, exsolution starts once the temperature at the CMB reaches a critical value corresponding to an MgO equilibrium concentration equal to that in the core. This is shown in Extended Data Fig. 5, and is highlighted for two models: the Mars-size impact[19] leaving behind a core containing 2.9 wt% MgO and a small fast-spinning impact[20] producing a core containing 2.1 wt% MgO (see Fig. 2b and Extended Data Fig. 3). The power produced by MgO exsolution is linked to the exsolution rate, and can be estimated from the energy release (Fig. 3 and Extended Data Fig. 8) to be between 5.5 TW wt% Gyr^{-1} and 7 TW wt% Gyr^{-1}. This estimate allows us to translate an exsolution rate (Extended Data Fig. 4d) into exsolution power, as shown in Extended Data Figs 5b and 8.

What is noteworthy is that the initial MgO core content does not directly affect exsolution power. The latter is a function of only the exsolution rate, which is itself a function of core cooling rate. Initial MgO content sets only the onset of exsolution, as shown in Extended Data Fig. 5. Of course, higher MgO contents in the core entail an earlier onset of exsolution, a longer duration for buoyancy-driven exsolution power and, hence, much higher total exsolution energies, as shown in Fig. 3. This dichotomy could be mitigated had we self-consistently included MgO exsolution in the thermal evolution model of the core. MgO exsolution power markedly drops with the onset of inner-core growth, as a consequence of the drop in core cooling rate. At the present day, MgO exsolution should still produce about 1 TW of power, much lower than the approximately 3 TW produced by inner-core growth and driving the geodynamo. However, before inner-core growth, exsolution power is always higher than about 3 TW, demonstrating that MgO exsolution can conceivably drive a geodynamo as early as around 1 Gyr after core formation, and until the onset of inner-core growth.

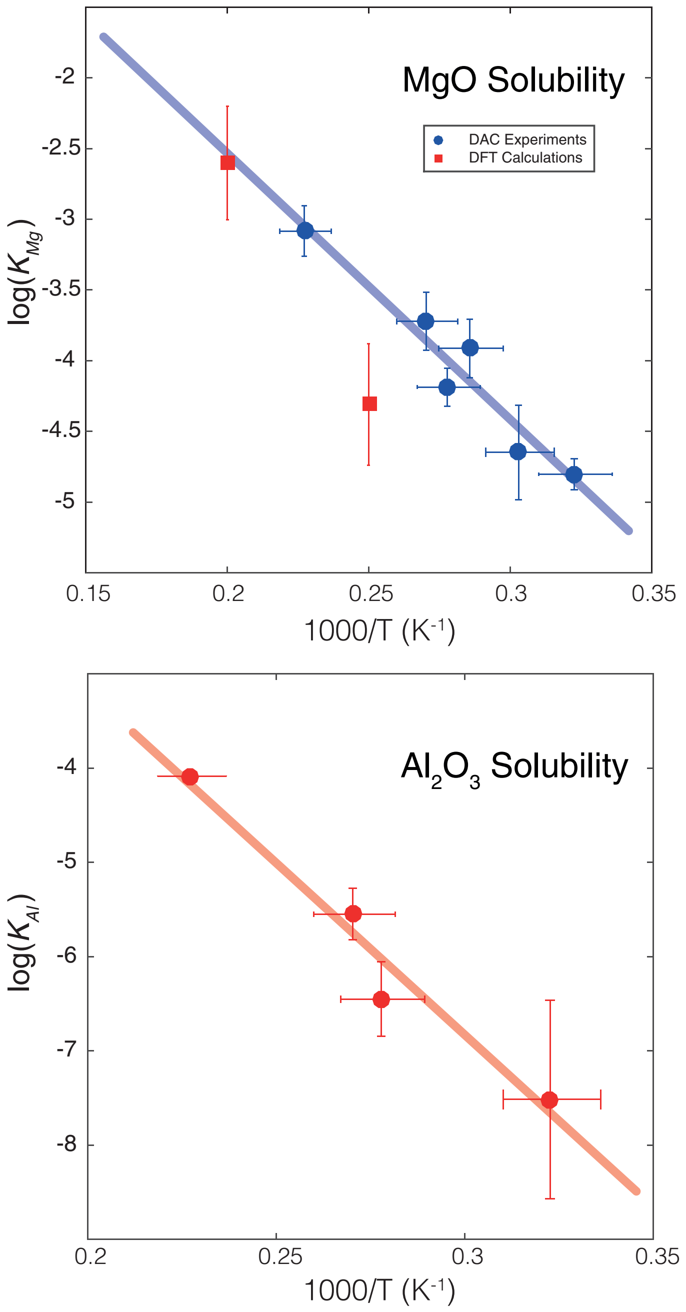
The geodynamo. Assuming an entirely bottom-driven present-day dynamo, corresponding to a CMB heat flow exactly at the adiabatic value (Q_{ad}) of 15 TW (refs 34,35), the convective power sustaining the geomagnetic field $P = \varepsilon Q_{\mathrm{ad}}$ is 3 TW, where $\varepsilon = 0.2$ is the thermodynamic efficiency of latent heat and light-element release at the inner-core boundary[22]. Power-based scaling laws of the magnetic intensity[36] then predict an internal magnetic field of about 1–4 mT, the higher estimate being in agreement with the observation of magnetic Alfvén waves in the core[37] coupled to length-of-day variations at periods close to 6 years (ref. 38).

Dynamo strength increases as buoyancy flux increases[39,40], so the MgO exsolution mechanism represents a potent driver of an early geodynamo[7]. Although a giant impact might cause thermal stratification in the core[6,41], the stabilizing thermal buoyancy will be completely overwhelmed by the compositional buoyancy associated with MgO exsolution.

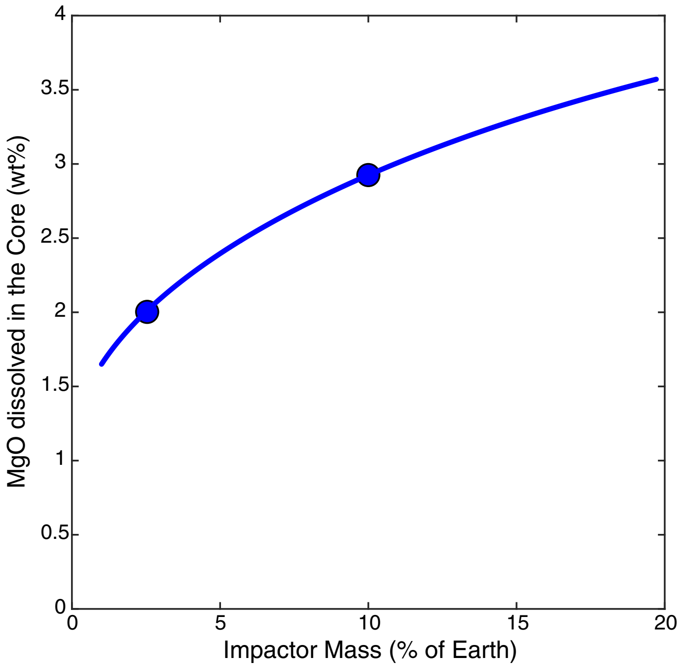
24. Kleine, T., Mezger, K., Munker, C., Palme, H. & Bischoff, A. Hf-182-W-182 isotope systematics of chondrites, eucrites, and martian meteorites: chronology of core formation and early mantle differentiation in Vesta and Mars. *Geochim. Cosmochim. Acta* **68,** 2935–2946 (2004).
25. Yin, Q. Z. *et al.* A short timescale for terrestrial planet formation from Hf-W chronometry of meteorites. *Nature* **418,** 949–952 (2002).
26. Murthy, V. R. Early differentiation of the Earth and the problem of mantle siderophile elements: a new approach. *Science* **253,** 303–306 (1991).
27. Li, J. & Agee, C. B. Geochemistry of mantle–core differentiation at high pressure. *Nature* **381,** 686–689 (1996).
28. Siebert, J., Badro, J., Antonangeli, D. & Ryerson, F. J. Metal-silicate partitioning of Ni and Co in a deep magma ocean. *Earth Planet. Sci. Lett.* **321–322,** 189–197 (2012).
29. Rudge, J. F., Kleine, T. & Bourdon, B. Broad bounds on Earth's accretion and core formation constrained by geochemical models. *Nat. Geosci.* **3,** 439–443 (2010).
30. Siebert, J., Badro, J., Antonangeli, D. & Ryerson, F. J. Terrestrial accretion under oxidizing conditions. *Science* **339,** 1194–1197 (2013).
31. Kleine, T., Mezger, K., Palme, H. & Munker, C. The W isotope evolution of the bulk silicate Earth: constraints on the timing and mechanisms of core formation and accretion. *Earth Planet. Sci. Lett.* **228,** 109–123 (2004).
32. Nakajima, M. & Stevenson, D. J. Dynamical mixing of planetary cores by giant impacts. *Lunar Planet. Sci. Conf.* **47,** 2053, http://www.hou.usra.edu/meetings/lpsc2016/pdf/2053.pdf (2016).
33. Fischer, R. A. *et al.* High pressure metal-silicate partitioning of Ni, Co, V, Cr, Si, and O. *Geochim. Cosmochim. Acta* **167,** 177–194 (2015).
34. Pozzo, M., Davies, C., Gubbins, D. & Alfè, D. Thermal and electrical conductivity of iron at Earth's core conditions. *Nature* **485,** 355–358 (2012).
35. de Koker, N., Steinle-Neumann, G. & Vlcek, V. Electrical resistivity and thermal conductivity of liquid Fe alloys at high P and T, and heat flux in Earth's core. *Proc. Natl Acad. Sci. USA* **109,** 4070–4073 (2012).
36. Christensen, U. R. A deep dynamo generating Mercury's magnetic field. *Nature* **444,** 1056–1058 (2006).
37. Gillet, N., Jault, D., Canet, E. & Fournier, A. Fast torsional waves and strong magnetic field within the Earth's core. *Nature* **465,** 74–77 (2010).
38. Buffett, B. A. Gravitational oscillations in the length of day. *Geophys. Res. Lett.* **23,** 2279–2282 (1996).
39. Olson, P. & Christensen, U. R. Dipole moment scaling for convection-driven planetary dynamos. *Earth Planet. Sci. Lett.* **250,** 561–571 (2006).
40. Christensen, U. R. & Aubert, J. Scaling properties of convection-driven dynamos in rotating spherical shells and application to planetary magnetic fields. *Geophys. J. Int.* **166,** 97–114 (2006).
41. Arkani-Hamed, J. & Olson, P. Giant impacts, core stratification, and failure of the Martian dynamo. *J. Geophys. Res. Solid Earth* **115,** E07012 (2010).



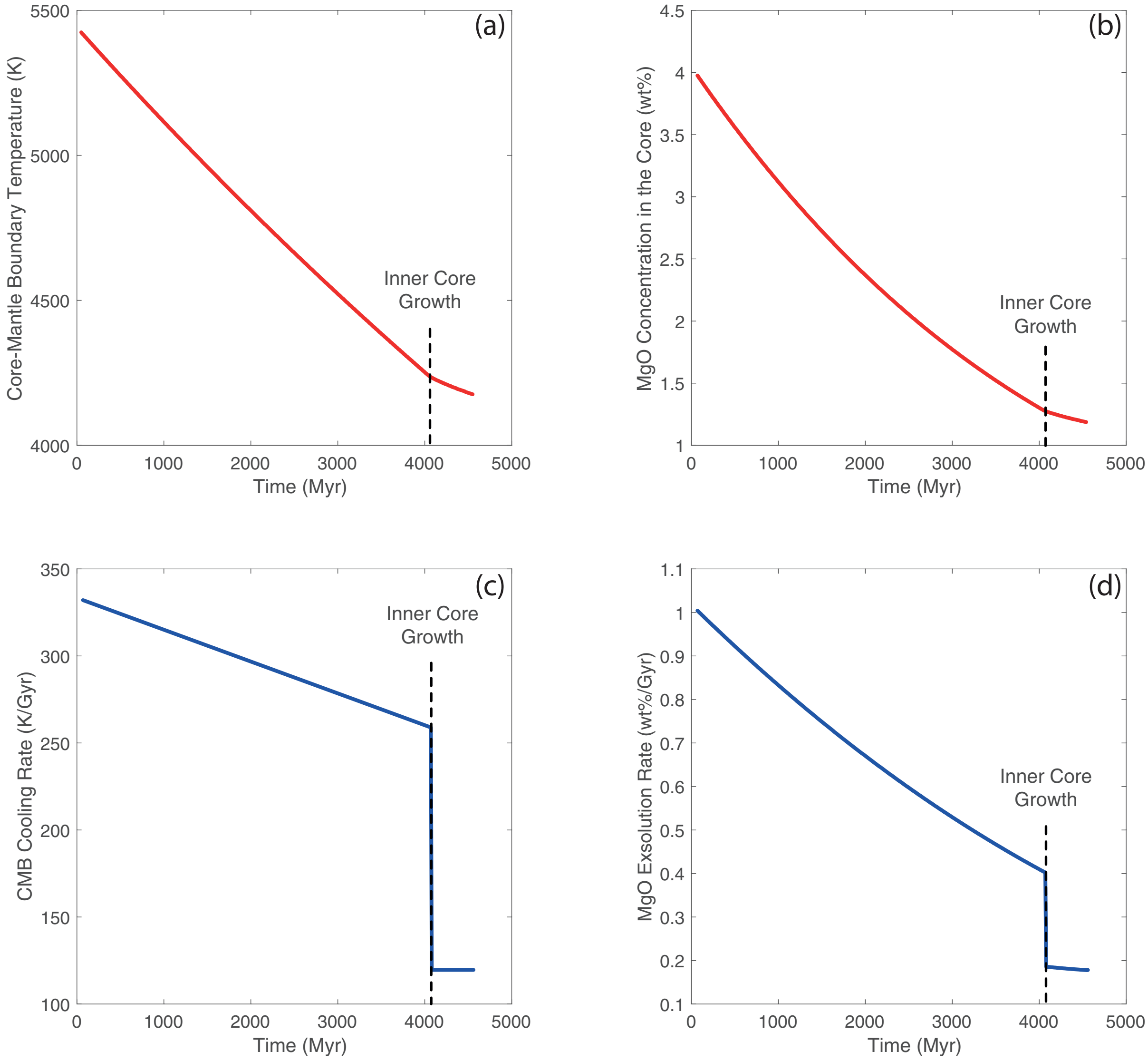
Extended Data Figure 1 | A fully molten metal-silicate sample recovered from the laser-heated diamond-anvil cell. A backscattered electron scanning electron microscopy image of a thin section recovered from a laser-heated diamond-anvil cell experiment. The section is excavated and lifted out from the centre of the heated region, then thinned down to 3 μm using a focused-ion-beam instrument. The metal and the silicate are both fully molten, as indicated by the coalesced metallic ball in the centre and the circular rim of silicate around it. This sample was compressed to 55 GPa and heated to 3,600 K for 60 s.



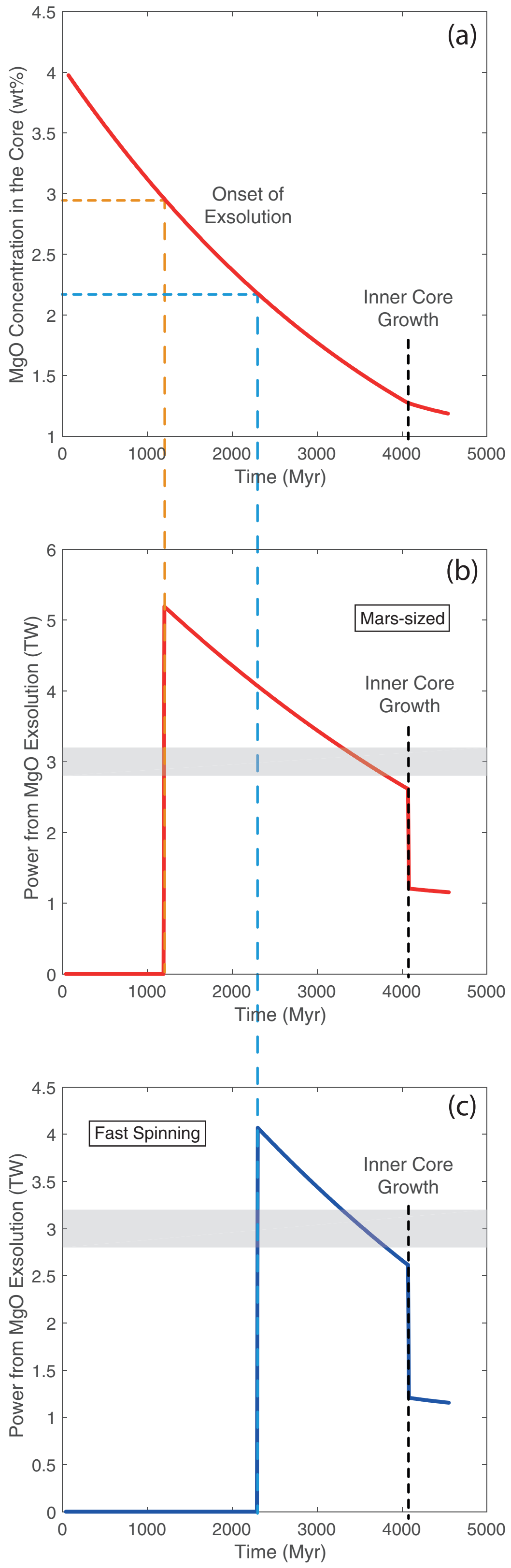
Extended Data Figure 2 | Magnesium and aluminium solubility in metallic iron melt at high pressure and temperature. Top, Equilibrium constant for MgO dissolution in molten iron (K_{Mg}) as a function of reciprocal temperature (1,000/T). The blue circles correspond to the experimental data (performed in a diamond-anvil cell, DAC; Extended Data Table 1) and the error bars to standard error; the red squares correspond to the low-temperature extrapolation of DFT calculations[17] and the error bars to standard error. The thick line corresponds to the least-squares linear fit to the experimental data (Fig. 1); it shows the agreement between the theoretical and experimental datasets, especially at high temperature where the theoretical dataset (which is extrapolated from higher temperatures) is the least influenced by extrapolation. Bottom, Equilibrium constant for Al_2O_3 dissolution (K_{Al}; see Methods) in molten iron as a function of reciprocal temperature. The circles correspond to the experimental data (Extended Data Table 1) and the error bars to standard error. The thick line corresponds to the least-squares linear fit to the data ($R^2 = 0.92$), and we find $\log(K_{\mathrm{Al}}) = 4.1(1.4) - 36{,}469(5{,}260)/T$ (equation (8)).



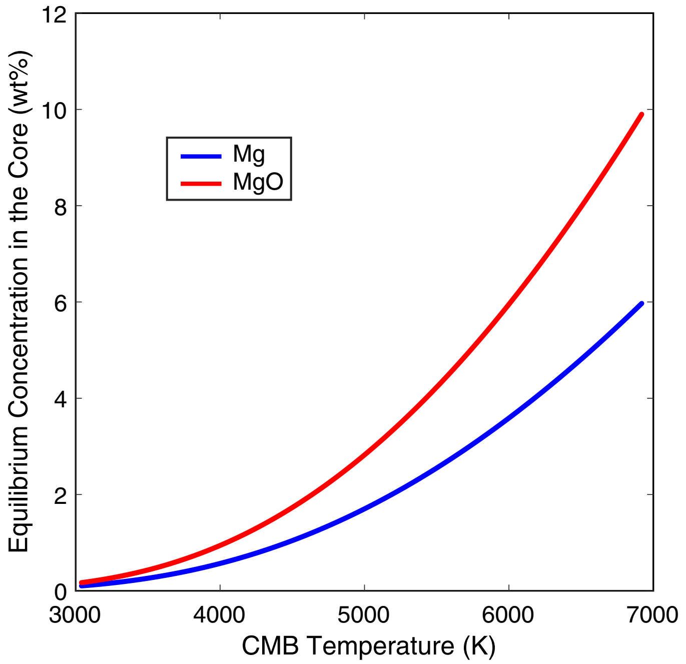
Extended Data Figure 3 | Total MgO dissolved in the core after the giant impact. A companion to Fig. 2, showing the sum of the MgO component dissolved in the core before the impact (0.8 wt%) and that brought by the HIC. The Mars-size impact[19] (10% of Earth's mass) and the fast-spinning impact[20] (2.5% of Earth's mass) are highlighted by circles, and provide 2.9 wt% and 2 wt% MgO to the core, respectively.



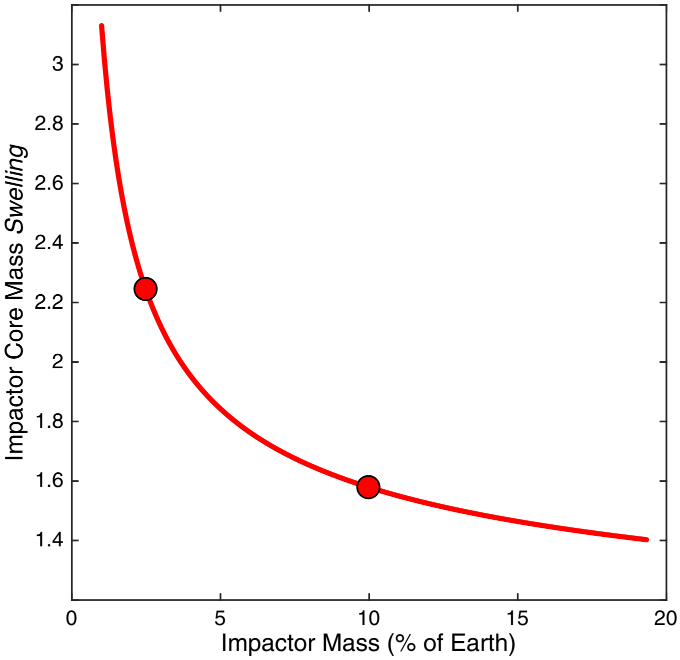
Extended Data Figure 4 | Thermal evolution of the core and MgO exsolution rate. a, **c**, Example CMB temperature evolution as a function of time (after Earth formation; Ga, billions of years ago), calculated using the same input parameters as in figure 4a of ref. 18 (**a**) and its derivative (**c**), which is the cooling rate. **b**, **d**, The associated MgO equilibrium concentration in the core (**b**), obtained by turning the temperature dependence in Fig. 1b into time dependence and its derivative (**d**), which is the exsolution rate. MgO will start exsolving from the core only when the MgO equilibrium concentration (**b**) drops below the MgO content in the core inherited from core formation and the giant impact. The core cooling rate and therefore the MgO exsolution rate drop markedly with the onset of inner-core growth. The core at the present day is still exsolving MgO, albeit at a much slower rate than that before inner-core growth.



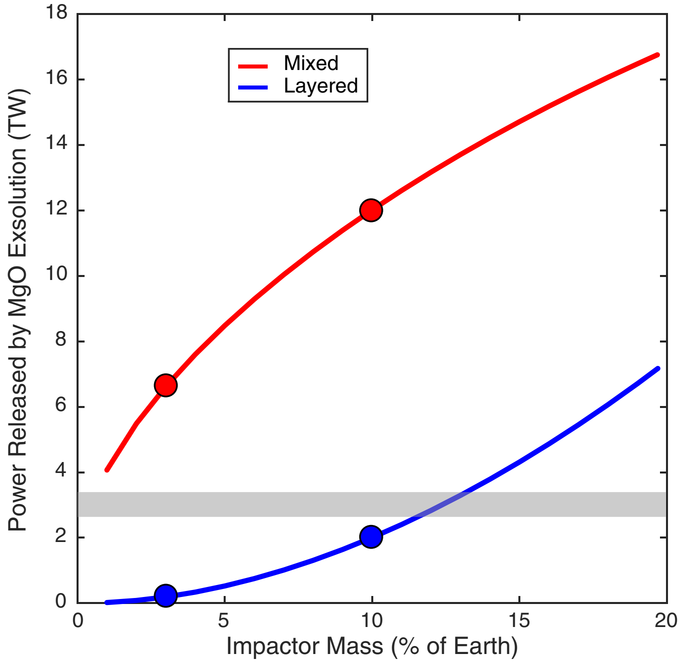
Extended Data Figure 5 | Onset of MgO exsolution and associated exsolution power for two typical models. a, The MgO equilibrium concentration in the core (same figure as Extended Data Fig. 4b), corresponding to our nominal CMB temperature evolution. The onset of MgO exsolution from the core occurs when the MgO equilibrium concentration drops below the MgO content in the core, which is reported here in two cases: 2.9 wt% for the Mars-sized impactor and 2.1 wt% for the fast-spinning impactor. For the thermal evolution model in Extended Data Fig. 4a, this onset is at 1.1 Gyr ago and 2.3 Gyr ago, respectively. **b**, **c**, Exsolution power for these two cases, which is proportional to the MgO exsolution rate plotted in Extended Data Fig. 4d. The power at a given time is independent of initial MgO content (as long as MgO is being exsolved). The initial MgO content affects only the onset of exsolution and therefore the duration of energy release. The power produced is in excess of 3 TW and is therefore sufficient to drive a dynamo by compositional buoyancy. The power drops markedly with the onset of inner-core growth, owing to the associated drop in the core cooling rate and the MgO exsolution rate. The core at the present day is still exsolving MgO and should produce about 1 TW of power, less than the power produced by inner-core growth.



Extended Data Figure 6 | Equilibrium Mg and MgO concentration in the core as a function of CMB temperature. This is obtained by rewriting $\log(K_{Mg}) = 1.23 - 18{,}816/T = 2\log(X_{Mg}) - \log(X_{MgO})$ as $\log(X_{Mg}) = [1.23 - 18{,}816/T + \log(X_{MgO})]/2$ with $X_{MgO} = 0.5$ (pyrolitic mantle). This curve (red for MgO, blue for Mg) allows us to determine the magnesium saturation in the core at a given temperature. This threshold is important to estimate: (i) the present-day MgO content of the core and, hence, the amount of MgO lost by exsolution over geologic time (Extended Data Fig. 4) and (ii) the temperature at which MgO exsolution started after core formation (Extended Data Fig. 5). For instance, for a core containing 2.9 wt% MgO (for a Mars-sized impact; see Extended Data Fig. 3), exsolution is not bound to occur until the temperature at the CMB cools below 5,030 K. Moreover, if the present-day CMB temperature is 4,100 K, then the MgO saturation in the present-day core is 1.1 wt%, so that the total amount of MgO that can be exsolved from the core is not the total initial MgO content, but that amount minus the present-day saturation value.



Extended Data Figure 7 | Chemical effect of equilibration of the impactor's core in Earth's magma ocean. Another companion to Fig. 2, showing the 'swelling' of the impactor core to form the hybridized impactor core (HIC). The HIC is larger than the impactor core because of the dissolved mantle components therein, which can represent up to two times its initial mass. This *y* axis shows the swelling factor, that is, the ratio of the mass of the HIC to that of the impactor core ($M_{\mathrm{HIC}}/M_{\mathrm{IC}}$). This swelling factor is equivalent to an effective dilution ratio. Small impactors interact with larger relative fractions of the magma ocean; therefore, they incorporate more mantle components per unit mass than do large impactors and so swell more. The HIC of a fast-spinning impactor[20] (2.5% of Earth's mass) is 2.2 times larger than the original impactor core, with 45% of its mass made up of initial impactor core material (iron) and the remaining 55% consisting of magma ocean components, as shown in Fig. 2a. The core of a Mars-sized impactor[19] (10% of Earth's mass) is 60% larger after equilibration with the magma ocean.



Extended Data Figure 8 | Power released by exsolution if it occurs over 1 Gyr. A companion to Fig. 3, showing how the gravitational energy released by exsolution is converted into average power, assuming a characteristic time of exsolution of 1 Gyr. The red curve corresponds to the energy released if the HIC fully mixes with Earth's core and the blue curve corresponds to the energy released if the HIC forms a layer on top of Earth's core. The grey horizontal band corresponds to 3 TW—the power driving the dynamo today—and thus provides a conservative estimate as to how much power is required to run a geodynamo by compositional buoyancy[22]. The Mars-size impact[19] (10% of Earth's mass) and fast-spinning impact[20] (2.5% of Earth's mass) are highlighted by circles. The blue curve represents a lower bound on the energy released in the case of layering of the HIC, because the layer contains so many mantle components that they would exsolve much faster, producing more power, albeit during a shorter period. By proportionality, this plot can be used to infer the power release for any characteristic exsolution time.

Extended Data Table 1 | Analyses of the Mg and Al concentrations in the metal and silicate phases of the experimental runs

Run	X1_2	X1_3	X1_4	X2_4	X4_2	X6_1
P (GPa)	71 (5)	35 (3)	50 (4)	74 (5)	55 (4)	43 (3)
T (K)	3500 (140)	3300 (130)	3700 (150)	4400 (180)	3600 (150)	3100 (130)
Mg (metal)	0.0042	0.0017	0.0088	0.0094	0.0053	0.0026
std err	0.0006	0.0004	0.0012	0.0011	0.0005	0.0002
MgO (silicate)	0.1446	0.1285	0.4081	0.1073	0.4324	0.4285
std err	0.0062	0.0034	0.0286	0.0072	0.0036	0.0070
log K_D	-3.9	-4.7	-3.7	-3.1	-4.2	-4.8
std err	0.19	0.35	0.21	0.18	0.13	0.10
Al (metal)			0.0018	0.0113	0.0008	0.0003
std err			0.0003	0.0002	0.0002	0.0002
$AlO_{1.5}$ (silicate)			0.0486	0.1651	0.0512	0.0511
std err			0.0036	0.0068	0.0012	0.0013
log K_D			-5.5	-4.1	-6.5	-7.4
std err			0.26	0.05	0.39	1.07

Experimental conditions (pressure in gigapascals, temperature in kelvin, uncertainties (standard error) in parentheses) and phase composition; all compositions are in molar fractions and standard errors are 1σ. The values for log(K_{Mg}) and log(K_{Al}) are plotted in Fig. 1 and Extended Data Fig. 2. Full chemical analyses of the samples are provided as Supplementary Data.

doi:10.1038/nature19093

Natural courtship song variation caused by an intronic retroelement in an ion channel gene

Yun Ding[1], Augusto Berrocal[1]†, Tomoko Morita[1], Kit D. Longden[1] & David L. Stern[1]

Animal species display enormous variation for innate behaviours, but little is known about how this diversity arose. Here, using an unbiased genetic approach, we map a courtship song difference between wild isolates of *Drosophila simulans* and *Drosophila mauritiana* to a 966 base pair region within the *slowpoke* (*slo*) locus, which encodes a calcium-activated potassium channel[1]. Using the reciprocal hemizygosity test[2], we confirm that *slo* is the causal locus and resolve the causal mutation to the evolutionarily recent insertion of a retroelement in a *slo* intron within *D. simulans*. Targeted deletion of this retroelement reverts the song phenotype and alters *slo* splicing. Like many ion channel genes, *slo* is expressed widely in the nervous system and influences a variety of behaviours[3,4]; *slo*-null males sing little song with severely disrupted features. By contrast, the natural variant of *slo* alters a specific component of courtship song, illustrating that regulatory evolution of a highly pleiotropic ion channel gene can cause modular changes in behaviour.

During courtship in *Drosophila* species, males vibrate their wings to produce a 'song' that attracts females[5]. Courtship song is relatively easy to quantify[6] and varies widely between species[7], making song an excellent system for genetic studies. However, despite decades of work[8–11], no causative loci for natural variation in courtship song have been identified definitively[12]. In particular, candidate gene approaches have failed to identify loci contributing to natural variation[13]. We have therefore taken an unbiased, whole-genome approach to identify loci underlying natural variation in courtship song.

Male flies in the *D. melanogaster* species subgroup, which includes the species studied here, produce courtship song that often contains two components: trains of continuous, approximately sinusoidal sound at a certain carrier frequency, called 'sine song', and a series of pulses separated by a characteristic interval, called 'pulse song'[5,7] (Fig. 1a and Supplementary Video 1). *D. simulans* and *D. mauritiana* diverged about 240 thousand years ago[14] and many features of their songs have changed. For example, the carrier frequency of sine song differs by 9.7 Hz between two wild-type isolates of these species, *sim5* and *mau29* (Fig. 1b and Supplementary Audios 1 and 2). We performed quantitative trait locus (QTL) mapping of courtship song traits between *sim5* and *mau29* using a high-throughput song phenotyping platform[6] and multiplexed shotgun genotyping[15]. The F1 hybrids and backcross progeny produced sine song with a frequency similar to *mau29*, indicating that the *mau29* allele(s) is largely dominant over the *sim5* allele(s) (Fig. 1b). In both backcrosses, we detected a single significant QTL at about 44.9 Mb on chromosome 3 (Fig. 1c) and the QTL explains most of the difference in sine frequency (Fig. 1d). We also identified one QTL for pulse song carrier frequency (Extended Data Fig. 1a, b) and two QTLs for inter-pulse interval (Extended Data Fig. 1c, d). The QTLs for these traits are located at different positions, indicating that different song features are genetically separable, consistent with the genetic modularity observed for other evolved behaviours[16,17].

To validate and fine-map the sine frequency QTL, we produced *D. mauritiana white* (*mauW*) strains with randomly inserted transposable elements carrying 3XP3::EYFP marker to facilitate targeted introgression. Using a marker located at 47.75 Mb on chromosome 3, we introgressed *D. mauritiana* DNA near the QTL into *sim5* (Fig. 2a). Of the many lines screened, several critical lines delimited the causal locus within an interval of about 1 Mb (44.9–45.9 Mb) that includes the

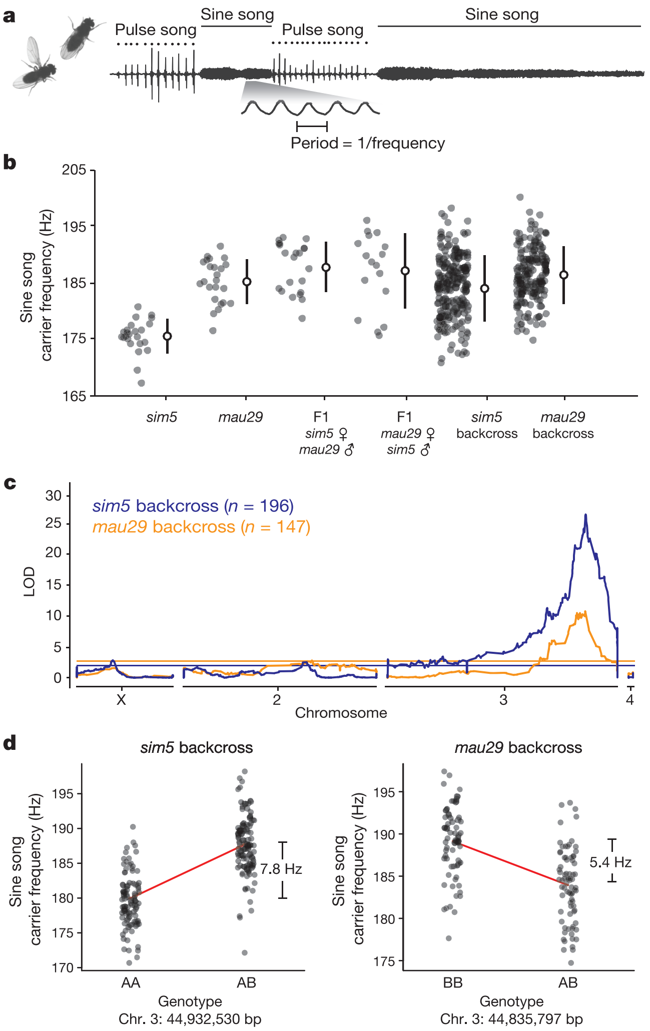


Figure 1 | QTL analysis of sine song frequency difference between *sim5* and *mau29*. a, Illustration of pulse and sine song. **b**, Sine song frequency in parental strains, F1 hybrids, and backcross males. Mean ± s.d. **c**, QTL map of *sim5* (blue) and *mau29* (orange) backcross. LOD, logarithm of the odds. Horizontal lines mark $P = 0.01$. **d**, Effect plots of the chromosome 3 QTL from each backcross. A, *sim5* allele; B, *mau29* allele. Red lines connect means for each genotype.

[1]Janelia Research Campus, Howard Hughes Medical Institute, Ashburn, Virginia 20147, USA. †Present address: Department of Molecular and Cell Biology, University of California, Berkeley, California 94720-3200, USA.

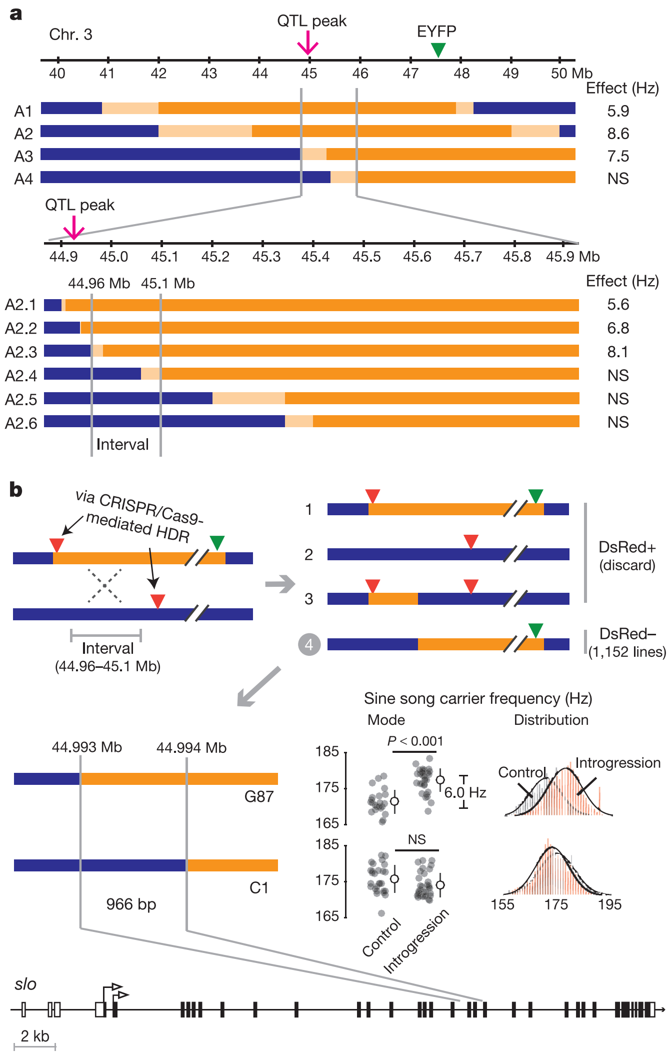


Figure 2 | Fine-scale mapping identifies *slo* as the candidate causal locus. a, Introgression of a *mauW* EYFP marker (green triangle) into a *sim5* background identifies a 140 kb region that causes a sine song frequency difference. Bar colour denotes species identity of DNA: blue, *sim5* DNA; orange, *mauW* DNA; light orange, breakpoint region. **b**, High-resolution mapping resolves the causal mutations to a 966 bp region of the *slo* locus. Upper panel, schematic of the targeted mapping strategy. Green and red triangles represent EYFP and DsRed markers. The products of either the absence (1, 2) or presence (3, 4) of recombination between DsRed markers are indicated. Middle panel, sine song frequency phenotypes (mean ± s.d.) of the two recombinant lines defining the minimal interval. *P* value by one-way ANOVA; NS, non-significant. Lower panel, diagram of *slo* gene structure showing minimal mapping interval. Black and open boxes indicate coding and non-coding exons, respectively, and open triangles mark the two alternative start codons. Genotyping and phenotyping data provided in Extended Data Fig. 2.

QTL peak (Fig. 2a and Extended Data Fig. 2a). This result validated our QTL map and suggests that *mau29* and *mauW* share the same genetic cause for higher sine frequency.

The *D. simulans* and *D. mauritiana* genomes differ at approximately one in every hundred base pairs and most differences are presumably irrelevant to song evolution. Therefore, unlike mapping of laboratory-induced mutations, we required high resolution to localize the causal nucleotides. To gain further resolution, we generated 500 additional introgression lines, defining a causal region of about 140 kb (44.96–45.1 Mb) (Fig. 2a and Extended Data Fig. 2a), which was still too large to identify a candidate gene. We therefore designed a targeted mapping strategy to resolve the precise location of the causal locus. We employed CRISPR/Cas9-mediated homology-directed repair

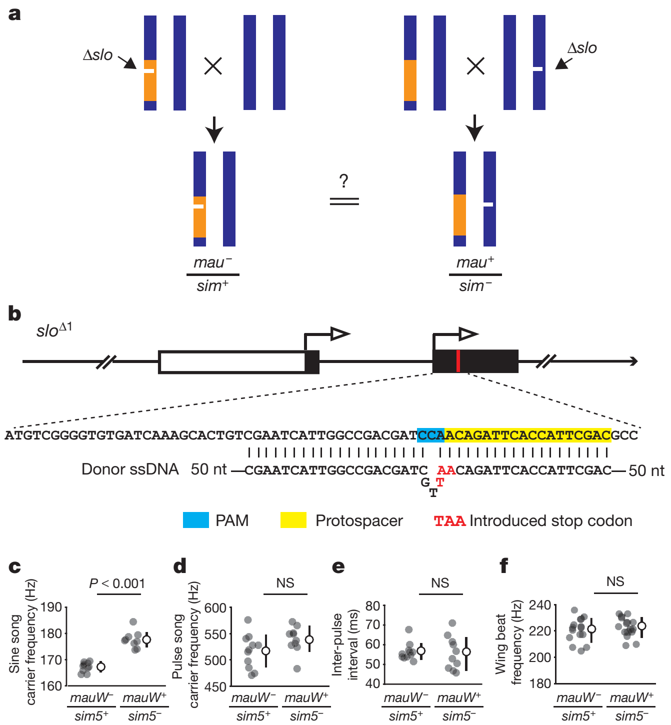


Figure 3 | Evolution of *slo* causes sine song frequency variation. a, Schematic of reciprocal hemizygosity test between *D. simulans* (*sim*) and *D. mauritiana* (*mau*) *slo* alleles in the genetic background of the introgression line A2.3. Blue and orange bars indicate *sim5* and *mauW* DNA, respectively. *Δslo*, *slo*-null allele. **b**, Design of *slo*-null allele $slo^{\Delta 1}$ induced by CRISPR/Cas9-mediated HDR using single-stranded DNA (ssDNA) as homology donor. PAM, protospacer adjacent motif; nt, nucelotides. **c**–**f**, Reciprocal hemizygosity test of behavioural phenotypes (mean ± s.d.) between *sim5* and *mauW*: sine song frequency (**c**); pulse song frequency (**d**); inter-pulse interval (**e**); and wing beat frequency (**f**). +, *slo* wild-type allele; −, *slo*-null allele $slo^{\Delta 1}$. *P* values by one-way ANOVA; NS, non-significant.

(HDR)[18] to insert 3XP3::DsRed markers at specific sites flanking the 140 kb interval (Extended Data Fig. 3) and identified recombinants in the interval by screening for eye colour (Fig. 2b). We generated 1,152 recombinants within a 170 kb region, providing mapping resolution of one recombination event per 148 bp, on average. All recombinants were genotyped and a subset were phenotyped when the mapped breakpoints could potentially further reduce the causal interval (Extended Data Fig. 2b, c). This effort identified a causal region of 966 bp within the *slo* locus (Fig. 2b), which encodes a calcium-activated potassium channel required to shape the excitability and firing pattern of neurons[1,19].

Although *slo* is a good candidate gene for courtship song variation, the causal mutation(s) in the mapped interval could, in principle, affect neighbouring genes and not *slo* itself. To directly test whether the evolved change(s) acts on the *slo* locus to alter sine frequency, we performed a reciprocal hemizygosity test, which is considered genetic proof for identifying causal genes underlying quantitative variation[2]. The test is implemented by generating null alleles in each of two strains and by crossing the mutant strains to the reciprocal non-mutated strains (Fig. 3a). The test, therefore, reveals the effects of alternative wild-type alleles in the same genetic background. We used CRISPR/Cas9-mediated HDR[18] to introduce a stop codon in the first coding exon shared by all *slo* splice isoforms in both the *sim5* and *mauW* alleles of the introgression line A2.3 (Fig. 3b and Extended Data Fig. 4). Hemizygous males carrying the *mauW slo* allele sang sine song at 10.0 Hz higher frequency than hemizygous males carrying the *sim5 slo* allele (Fig. 3c). Therefore, variation acting on *slo* causes a sine frequency difference between the two strains.

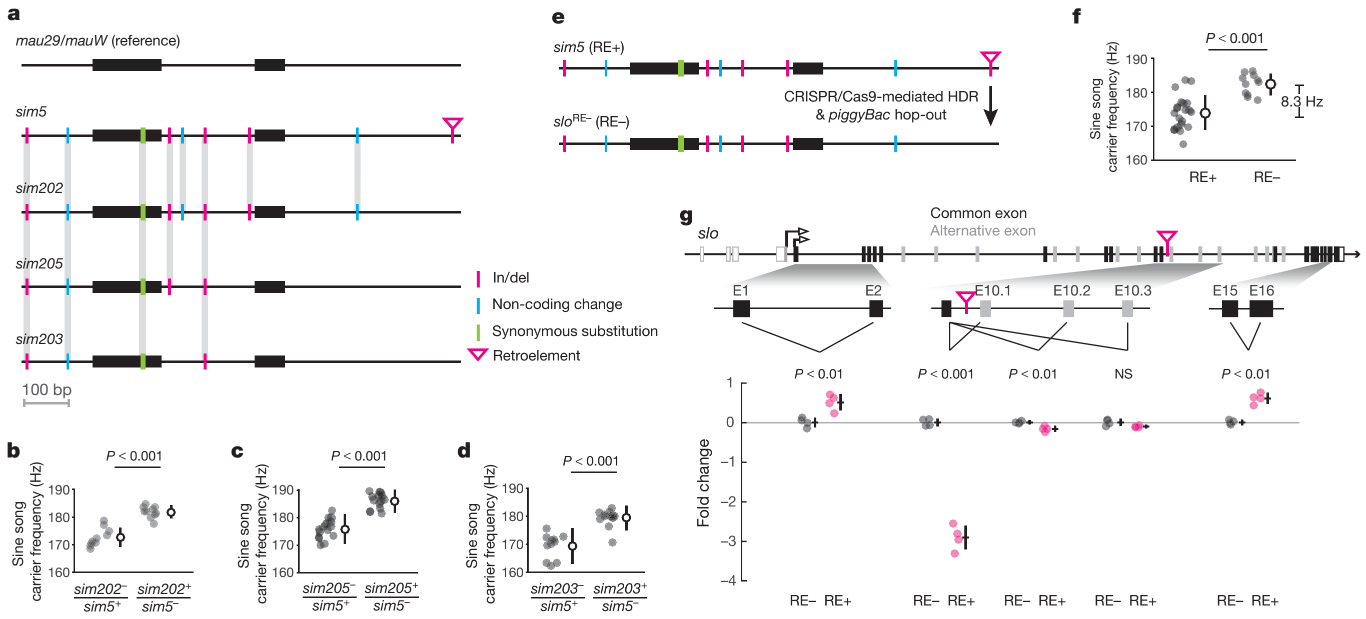


Figure 4 | Intronic insertion of a retroelement at the *slo* locus is the causal mutation. a, Candidate mutations in *D. simulans* strains *sim5*, *sim202*, *sim205*, and *sim203*, using *mau29* and *mauW* sequences as reference. Shared sites are highlighted by grey bars. **b**–**d**, Reciprocal hemizygosity test of sine frequency (mean ± s.d.) between *sim5* and *sim202* (**b**), between *sim5* and *sim205* (**c**), and between *sim5* and *sim203* (**d**). +, *slo* wild-type allele; −, *slo*-null allele *slo*$^{\Delta 2}$. *P* values by one-way ANOVA. **e**, Targeted deletion of the retroelement insertion in *sim5* to generate *slo*$^{\mathrm{RE}-}$. Experimental details provided in Extended Data Fig. 8. **f**, Comparison of sine song frequency (mean ± s.d.) between sim5 (RE+) and *slo*$^{\mathrm{RE}-}$ (RE−). *P* value by one-way ANOVA. **g**, Expression differences of five *slo* exon junctions between *sim5* (RE+, magenta) and *slo*$^{\mathrm{RE}-}$ (RE−, grey), assayed by quantitative reverse-transcription PCR (RT-qPCR). The alternative exons E10.1, E10.2 and E10.3 are mutually exclusive in full-length transcripts. Each exon junction was assayed in four biological replicates with two technical replicates. Mean ± s.d. *P* values by two-sided *t*-test; NS, non-significant.

The *slo* gene is expressed broadly in the fly nervous system[3] (Extended Data Fig. 5a, c, d) and influences many locomotor behaviours[4]. We found that *slo*-null males of *sim5* produced little song with disrupted pulse and sine events (Extended Data Fig. 5b, e–h). In contrast, in the reciprocal hemizygosity test, the evolved allele of *slo* altered only sine frequency, but not pulse song frequency (Fig. 3d), inter-pulse interval (Fig. 3e), wing beat frequency in tethered flight (Fig. 3f), or any other song traits we measured (Extended Data Fig. 6). Therefore, while *slo* has pleiotropic roles, the natural variant of *slo* alters a specific component of song.

There are ten candidate differences in the minimal mapping interval (Fig. 4a): two synonymous coding changes, three non-coding single nucleotide changes, four small non-coding insertions/deletions (in/del), and one 6.7 kb intronic retroelement. To identify the causal mutation(s), we exploited natural variation. We analysed songs from 12 *D. simulans* and 12 *D. mauritiana* wild-type isolates. On average, the sine song frequency of *D. simulans* is 7.8 Hz lower than *D. mauritiana*, although each species contains extensive variation for sine frequency (Extended Data Fig. 7). We sequenced the 966 bp minimal interval in all these strains and found that many of the candidate differences are polymorphic within species, potentially allowing reciprocal hemizygosity testing to narrow down the causal mutation(s).

We generated new *slo*-null alleles in three *D. simulans* strains, *sim202*, *sim205*, and *sim203* (Extended Data Fig. 5b), which harbour nine, six, and five of the ten candidate differences, respectively (Fig. 4a). We performed reciprocal hemizygosity tests between *sim5* and each strain. If the causal mutation is shared by two *D. simulans* strains, then we expect no sine frequency difference between the reciprocal hemizygotes; if the causal mutation is specific to *sim5*, then we expect to observe a difference. For each comparison, the hemizygote with a single *sim5* copy produced song with a significantly lower sine frequency (Fig. 4b–d). The 6.7 kb retroelement insertion is the only polymorphism that differs between *sim5* and all three other strains, making it the best candidate.

We then directly tested the effect of the retroelement by targeted deletion. We first replaced the retroelement with a 3XP3::DsRed marker via CRISPR/Cas9-mediated HDR and then removed the marker (Fig. 4e and Extended Data Fig. 8). The resultant flies, *slo*$^{\mathrm{RE}-}$, sang sine song at 8.3 Hz higher frequency than wild-type *sim5* (Fig. 4f). Therefore, the intronic retroelement in the *slo* locus is the causal mutation.

Among the 12 surveyed *D. simulans* strains, this intronic retroelement insertion was detected only in *sim5*. Therefore, it represents a newly derived, rare variant within *D. simulans*. The retroelement resembles a retrovirus (Extended Data Fig. 9a), although it is distinct from any previously characterized retroelement clades (Extended Data Fig. 10). We therefore named it *Shellder*. We identified many polymorphic putative *Shellder* insertions in wild-type strains of *D. simulans* and *D. mauritiana* using TagMap[20] (Extended Data Fig. 9b), suggesting that *Shellder* is probably propagating actively in *Drosophila* populations.

Like many ion channel genes, *slo* exhibits complex patterns of tissue-specific transcription[21] and alternative splicing[22], potentially producing channels with distinctive properties that could be exploited during evolution. Using the *Shellder* excision allele, we found that the *Shellder* insertion causes a 2.9-fold decrease specifically in the usage of its flanking exon junction but also a slight increase in the overall expression level of *slo* (Fig. 4g). Thus, the insertion does not appear to cause mRNA decay. Considering that lower *slo* level is associated with lower sine frequency (Extended Data Fig. 5i), the lower frequency phenotype caused by the *Shellder* insertion most likely results from the splicing changes of *slo*.

Previous studies have identified genetic variation influencing behaviour[23–28], including how variation in sensory systems alters the probability of particular behaviours[23–26]. Our study is the first, to our knowledge, to identify the genetic cause for variation in a motor pattern. Our study illustrates that specific behaviour changes can result from certain kinds of regulatory changes in a highly pleiotropic ion channel, in this case most likely through modifications of alternative splicing.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 21 October 2015; accepted 8 July 2016.

Published online 10 August; corrected online 17 August 2016 (see full-text HTML version for details).

1. Atkinson, N. S., Robertson, G. A. & Ganetzky, B. A component of calcium-activated potassium channels encoded by the *Drosophila slo* locus. *Science* **253,** 551–555 (1991).
2. Stern, D. L. Identification of loci that cause phenotypic variation in diverse species with the reciprocal hemizygosity test. *Trends Genet.* **30,** 547–554 (2014).
3. Becker, M. N., Brenner, R. & Atkinson, N. S. Tissue-specific expression of a *Drosophila* calcium-activated potassium channel. *J. Neurosci.* **15,** 6250–6259 (1995).
4. Atkinson, N. S. *et al.* Molecular separation of two behavioral phenotypes by a mutation affecting the promoters of a Ca-activated K channel. *J. Neurosci.* **20,** 2988–2993 (2000).
5. Bennet-Clark, H. C. & Ewing, A. W. The courtship songs of *Drosophila*. *Behaviour* **31,** 288–301 (1968).
6. Arthur, B. J., Sunayama-Morita, T., Coen, P., Murthy, M. & Stern, D. L. Multi-channel acoustic recording and automated analysis of *Drosophila* courtship songs. *BMC Biol.* **11,** 11 (2013).
7. Greenspan, R. J. & Ferveur, J. F. Courtship in *Drosophila*. *Annu. Rev. Genet.* **34,** 205–232 (2000).
8. Huttunen, S., Aspi, J., Hoikkala, A. & Schlötterer, C. QTL analysis of variation in male courtship song characters in *Drosophila virilis*. *Heredity* **92,** 263–269 (2004).
9. Gleason, J. M., Nuzhdin, S. V. & Ritchie, M. G. Quantitative trait loci affecting a courtship signal in *Drosophila melanogaster. Heredity* **89,** 1–6 (2002).
10. Turner, T. L., Miller, P. M. & Cochrane, V. A. Combining genome-wide methods to investigate the genetic complexity of courtship song variation in *Drosophila melanogaster*. *Mol. Biol. Evol.* **30,** 2113–2120 (2013).
11. Williams, M. A., Blouin, A. G. & Noor, M. A. F. Courtship songs of *Drosophila pseudoobscura* and *D. persimilis*. II. Genetics of species differences. *Heredity* **86,** 68–77 (2001).
12. Stern, D. L. Reported *Drosophila* courtship song rhythms are artifacts of data analysis. *BMC Biol.* **12,** 38 (2014).
13. Cande, J., Stern, D. L., Morita, T., Prud'homme, B. & Gompel, N. Looking under the lamp post: neither *fruitless* nor *doublesex* has evolved to generate divergent male courtship in *Drosophila*. *Cell Reports* **8,** 363–370 (2014).
14. Garrigan, D. *et al.* Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res.* **22,** 1499–1511 (2012).
15. Andolfatto, P. *et al.* Multiplexed shotgun genotyping for rapid and efficient genetic mapping. *Genome Res.* **21,** 610–617 (2011).
16. Greenwood, A. K., Wark, A. R., Yoshida, K. & Peichel, C. L. Genetic and neural modularity underlie the evolution of schooling behavior in threespine sticklebacks. *Curr. Biol.* **23,** 1884–1888 (2013).
17. Weber, J. N., Peterson, B. K. & Hoekstra, H. E. Discrete genetic modules are responsible for complex burrow evolution in *Peromyscus* mice. *Nature* **493,** 402–405 (2013).
18. Bassett, A. R. & Liu, J.-L. CRISPR/Cas9 and genome editing in *Drosophila*. *J. Genet. Genomics* **41,** 7–19 (2014).
19. Vergara, C., Latorre, R., Marrion, N. V. & Adelman, J. P. Calcium-activated potassium channels. *Curr. Opin. Neurobiol.* **8,** 321–329 (1998).
20. Stern, D. L. Tagmentation-based mapping (TagMap) of mobile DNA genomic insertion sites. Preprint at http://dx.doi.org/10.1101/037762 (2016).
21. Brenner, R., Thomas, T. O., Becker, M. N. & Atkinson, N. S. Tissue-specific expression of a Ca^{2+}-activated K^{+} channel is controlled by multiple upstream regulatory elements. *J. Neurosci.* **16,** 1827–1835 (1996).
22. Yu, J. Y., Upadhyaya, A. B. & Atkinson, N. S. Tissue-specific alternative splicing of BK channel transcripts in *Drosophila*. *Genes Brain Behav.* **5,** 329–339 (2006).
23. Bendesky, A., Tsunozaki, M., Rockman, M. V., Kruglyak, L. & Bargmann, C. I. Catecholamine receptor polymorphisms affect decision-making in *C. elegans*. *Nature* **472,** 313–318 (2011).
24. Bendesky, A. *et al.* Long-range regulatory polymorphisms affecting a GABA receptor constitute a quantitative trait locus (QTL) for social behavior in *Caenorhabditis elegans*. *PLoS Genet.* **8,** e1003157 (2012).
25. de Bono, M. & Bargmann, C. I. Natural variation in a neuropeptide Y receptor homolog modifies social behavior and food response in *C. elegans*. *Cell* **94,** 679–689 (1998).
26. McGrath, P. T. *et al.* Quantitative mapping of a digenic behavioral trait implicates globin variation in *C. elegans* sensory behaviors. *Neuron* **61,** 692–699 (2009).
27. Osborne, K. A. *et al.* Natural behavior polymorphism due to a cGMP-dependent protein kinase of *Drosophila*. *Science* **277,** 834–836 (1997).
28. Yalcin, B. *et al.* Genetic dissection of a behavioral quantitative trait locus shows that *Rgs2* modulates anxiety in mice. *Nat. Genet.* **36,** 1197–1202 (2004).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank C. Llorens for advice on retroelement classification and B. Dickson, N. Frankel, T. Shirangi, N. Spruston and S. Sternson for advice on the manuscript.

Author Contributions Y.D. and D.L.S. designed the experiments. A.B. performed QTL analysis, with help from T.M. Y.D. and K.D.L. assayed wing beat frequency. D.L.S. performed TagMap analysis. Y.D. performed all other experiments and analysis. Y.D. and D.L.S. wrote the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to D.L.S. (sternd@janelia.hhmi.org).

Reviewer Information *Nature* thanks B. Prud'homme, T. Turner and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment, but all statistical analysis was performed automatically by our analysis pipeline.

Strains and behavioural assays. Fly strains used are summarized in Supplementary Table 1. For all behavioural assays, the flies were reared in standard laboratory conditions. Courtship song was recorded as described previously[6]. For each comparison reported in this study, the flies were recorded simultaneously, and if applicable, collected from the same vials. Song parameters were estimated as the mode of song events across 30-min recordings. Wing beat frequency was measured on tethered male flies given a bar fixation task during flight, by optically tracking wing movements with Wingbeat Analyzer (JFI Electronics Laboratory, University of Chicago)[29].

Behaviour data analysis and statistics. Song data was segmented[6] and analysed (http://www.github.com/dstern/BatchSongAnalysis) without human intervention. For each test, our target sample size was $n = 12$ per genotype, with individuals selected haphazardly, for power of 0.8 to detect 1 s.d. difference between treatments at $P < 0.05$. Outliers were systematically excluded in our song analysis pipeline using the Grubbs test with $\alpha = 0.05$ (http://www.mathworks.com/matlabcentral/fileexchange/3961-deleteoutliers). *P* values for ANOVAs were estimated with 10,000 permutations (http://www.mathworks.com/matlabcentral/fileexchange/44307-randanova1). All critical experiments were replicated at least once with a similar sample size.

QTL mapping. QTL mapping employed 210 *sim5* and 180 *mau29* backcross progeny, which were processed into a single multiplexed shotgun genotyping (MSG) library[15]. Parental genomes were generated by updating the *D. simulans* r2.0.1 genome (http://www.flybase.org/static_pages/feature/previous/articles/2015_02/Dsim_r2.01.html) with HiSeq reads from each strain (SRA accession: SRP076910). Genotypes were estimated with MSG software (http://www.github.com/JaneliaSciComp/msg). Posterior probabilities of ancestry were thinned using pull_thin (http://www.github.com/dstern/pull_thin) and imported into R-QTL[30] using read_cross_msg (http://www.github.com/dstern/read_cross_msg). Genome scans with a single QTL model were performed using Haley–Knott regression[30] and *P* values were estimated with 1,000 permutations.

Fine-scale introgression mapping and high-resolution recombination mapping. Genetic mapping was performed in three phases. First, to develop a visible marker linked to the QTL, we screened a collection of *mauW* strains that had been transformed with a *piggyBac* transposable element carrying 3XP3::EYFP, which drives yellow fluorescent protein expression in the eyes. The strain *mauW446*, which carried a *piggyBac* insertion within 3 Mb of the QTL peak, was backcrossed to *sim5* for five generations. The genetic background of the resulting introgression lines was checked using whole genome sequencing. Second, the introgression line A2, which produces *D. mauritiana*-like sine song and does not contain any *D. mauritiana* DNA outside of the target region, was backcrossed to *sim5* for three to five generations for further introgression. Third, we inserted two 3XP3::DsRed markers into the introgression line A2.3: one on the left side of the interval on the EYFP-marked introgression chromosome (44.95 Mb), and the other on the right side on the *sim5* chromosome (45.12 Mb). Female flies carrying both DsRed markers were crossed to *sim5* males and male progeny without a DsRed marker were identified as recombinants. All these lines were maintained through the male lineage, which does not experience recombination. Recombination breakpoints were mapped using molecular markers (Supplementary Table 2) and sequencing.

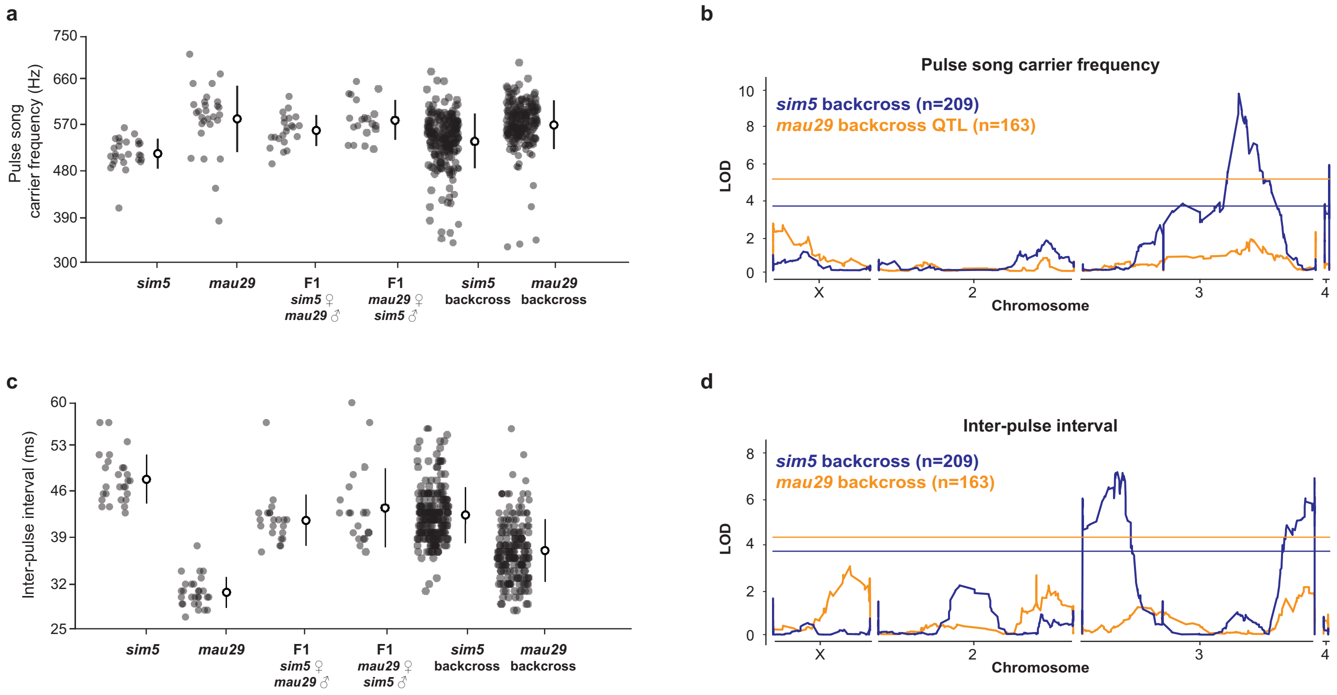
CRISPR/Cas9 genome editing. For all CRISPR/Cas9-mediated HDR, guide RNAs (gRNAs), donor DNA, *in vitro* transcribed *D. melanogaster* codon optimized Cas9 mRNA, and a Dicer-substrate short interfering RNA (DsiRNA) targeted against *lig4* (Sequence1: rUrCrCrUrGrCrArGrCrUrGrArUrGrCrUrUrGrCrUrGrUrGrUrCrGrU; Sequence 2: rGrArCrArCrArGrCrArArGrCrArUrCrArGrCrUrGrCrArGG A, synthesized by IDT) to inhibit non-homologous end joining[18] were co-injected into the embryos. All the injections were performed by Rainbow Transgenic Flies using the following concentrations: 0.2 μg/μl gRNA source, 0.5 μg/μl donor DNA, 0.1 μg/μl Cas9 mRNA, and 0.1 μg/μl *lig4* DsiRNA. The germline transmission rates are provided in Supplementary Table 3.

RT-qPCR. Total RNA was extracted from 5–7 day old males and converted to cDNA template after DNase I treatment. Real-time PCR was performed using *ACTB* as an internal control. The primer sequences are provided in Supplementary Table 4.

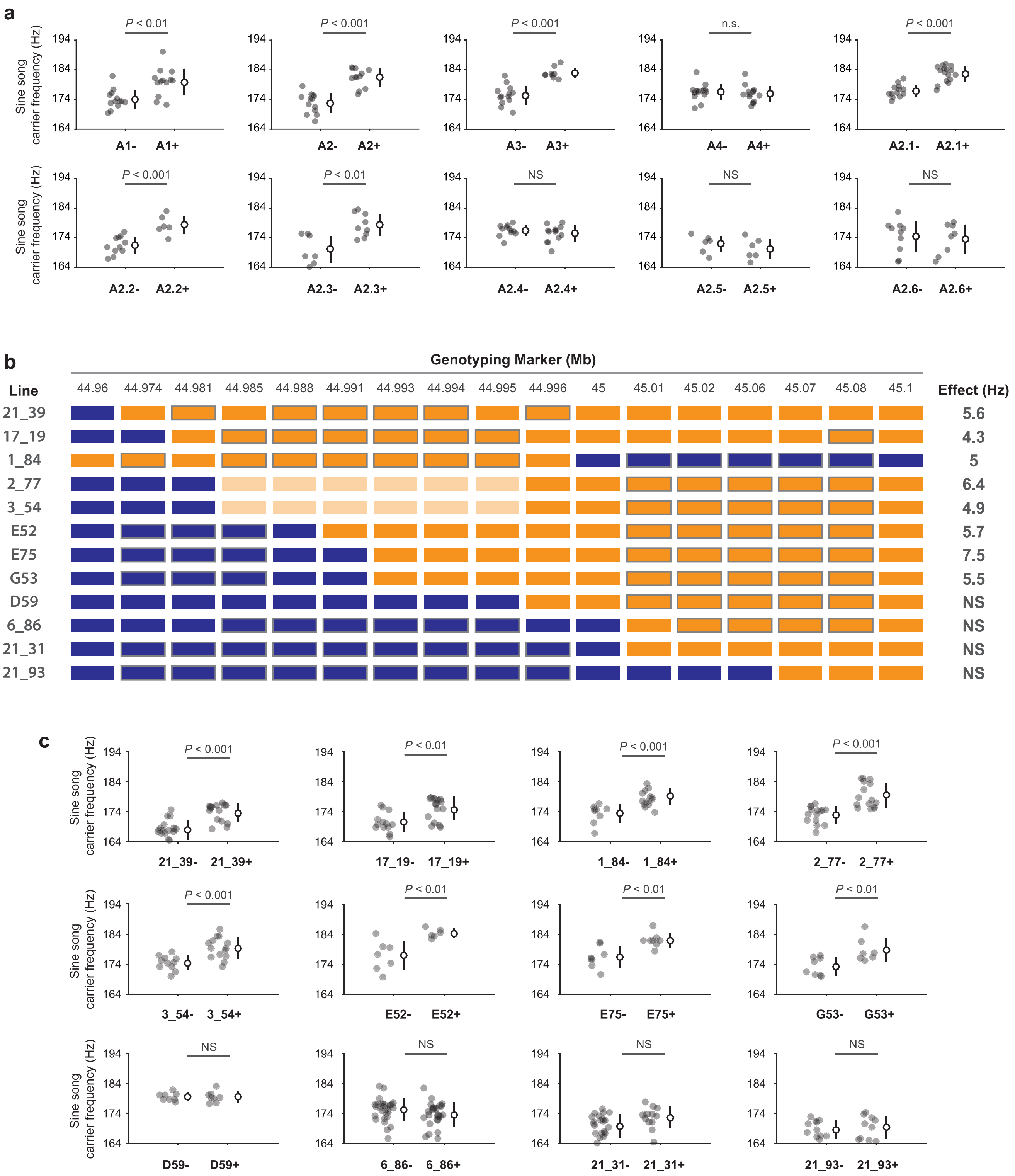
***Shellder* analysis.** The name comes from the Pokémon character Shellder, who can attach to the character Slowpoke and cause Slowpoke to evolve into a new form. The sequence annotation, phylogenetic analysis, and TaqMap identification of *Shellder* (GenBank KX196449) insertion sites are described in Supplementary Methods.

29. Gotz, K. G. Course-control, metabolism and wing interference during ultralong tethered flight in *Drosophila melanogaster*. *J. Exp. Biol.* **128,** 35–46 (1987).

30. Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19,** 889–890 (2003).

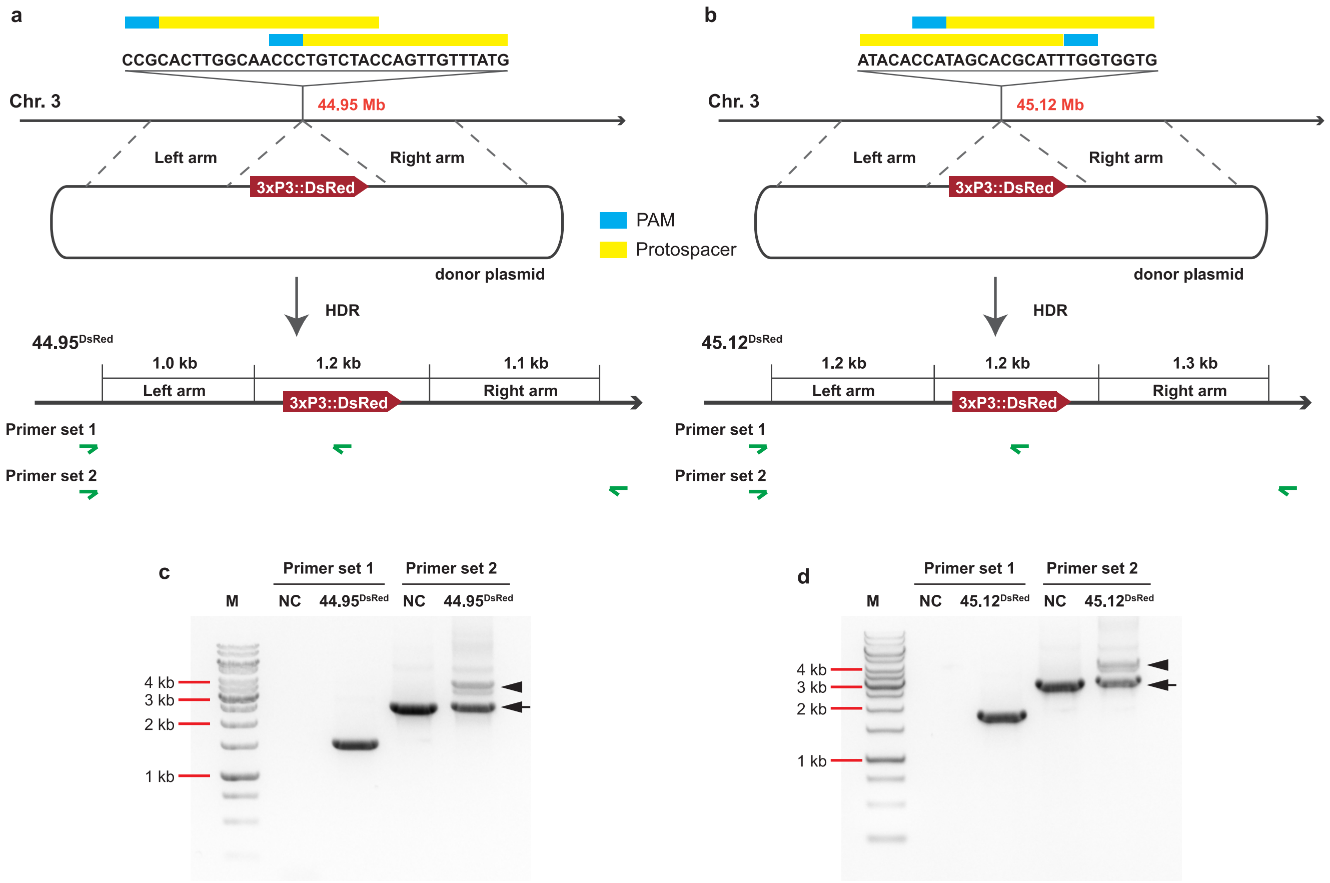


Extended Data Figure 1 | QTL analysis of pulse song carrier frequency and inter-pulse interval. a, Pulse song carrier frequency (mean ± s.d.) in parental strains, F1 hybrids, and backcross males. **b**, QTL map of pulse song frequency. LOD, logarithm of the odds. Horizontal lines mark $P = 0.01$. A single QTL on chromosome 3 at 34,919,124 bp was identified in the *sim5* backcross (blue). No significant QTL were detected in the *mau29* backcross (orange). **c**, Inter-pulse interval (mean ± s.d.) in parental strains, F1 hybrids, and backcross males. **d**, QTL map of inter-pulse interval. Two QTLs on chromosome 3 at 7,902,342 bp and 52,144,317 bp were identified in the *sim5* backcross (blue). No significant QTL were detected in the *mau29* backcross (orange).

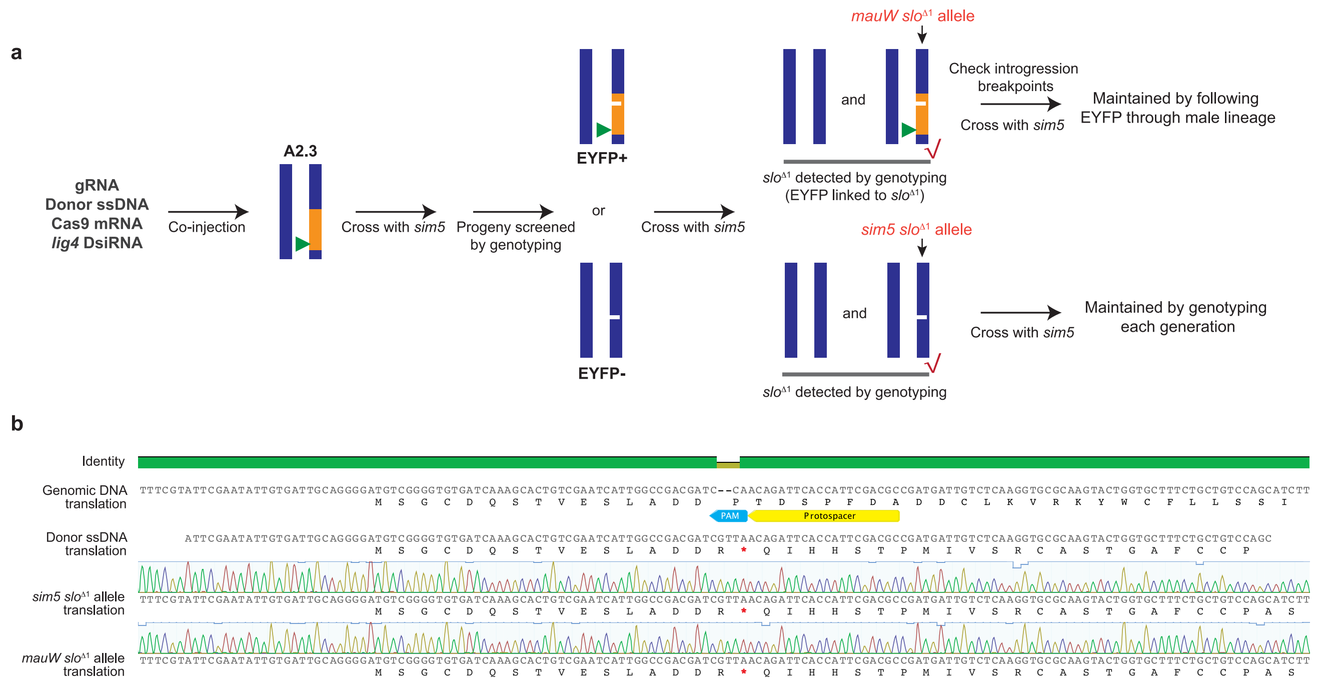


Line	44.96	44.974	44.981	44.985	44.988	44.991	44.993	44.994	44.995	44.996	45	45.01	45.02	45.06	45.07	45.08	45.1	Effect (Hz)
21_39	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	5.6
17_19	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	4.3
1_84	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟦	🟦	🟦	🟦	🟦	🟦	🟦	5
2_77	🟦	🟦	🟦	▨	▨	▨	▨	▨	▨	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	6.4
3_54	🟦	🟦	🟦	▨	▨	▨	▨	▨	▨	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	4.9
E52	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	5.7
E75	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	7.5
G53	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	5.5
D59	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	NS
6_86	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	NS
21_31	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	🟧	🟧	🟧	NS
21_93	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟧	🟧	🟧	NS

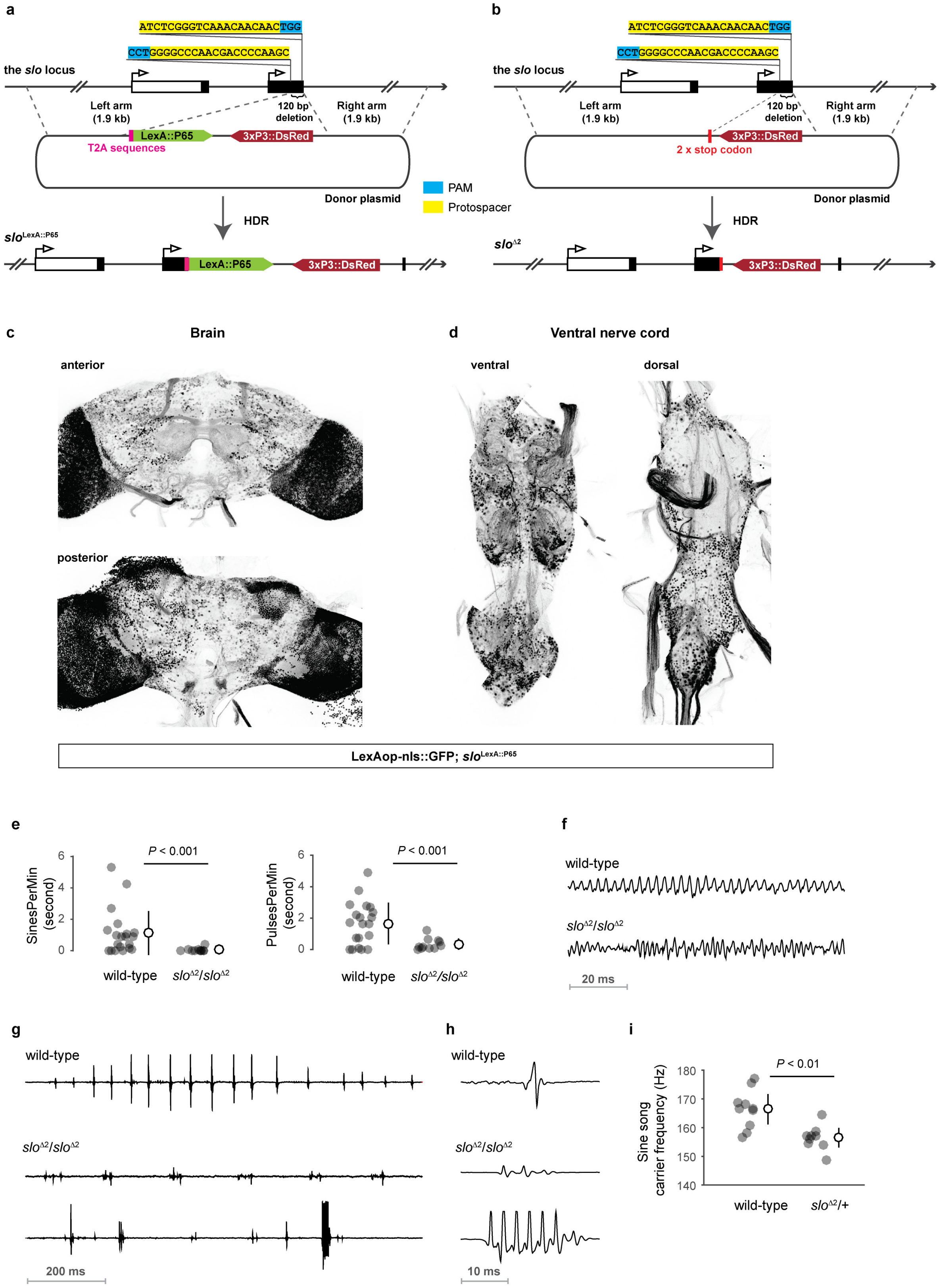
Extended Data Figure 2 | Details of fine-scale mapping. a, Sine song frequency of the introgression lines shown in Fig. 2a. **b**, Genotype and phenotype data of informative recombinant lines. The genotyping markers are listed on the top, according to their physical locations on chromosome 3. For each recombinant line, the genotyping results are represented using coloured bars: blue, homozygote for *sim5*; orange, heterozygote for *sim5* and *mauW*; and light orange, unknown. For boxed bars, the genotypes were assigned assuming no additional recombination events in this region. The sine song frequency effect is summarized on the right. Line 1_84 was recovered by selecting flies with stronger DsRed and without EYFP. (This strategy was abandoned later due to the challenge of distinguishing two copies of DsRed from one. Please refer to Fig. 2b.) **c**, Sine song frequency phenotypes of the recombinant lines in panel **b**. Comparison of sine song frequency (mean ± s.d.) was made between heterozygous introgression males (+) and their pure *sim5* sibling brothers (−) using one-way ANOVA. NS, non-significant.



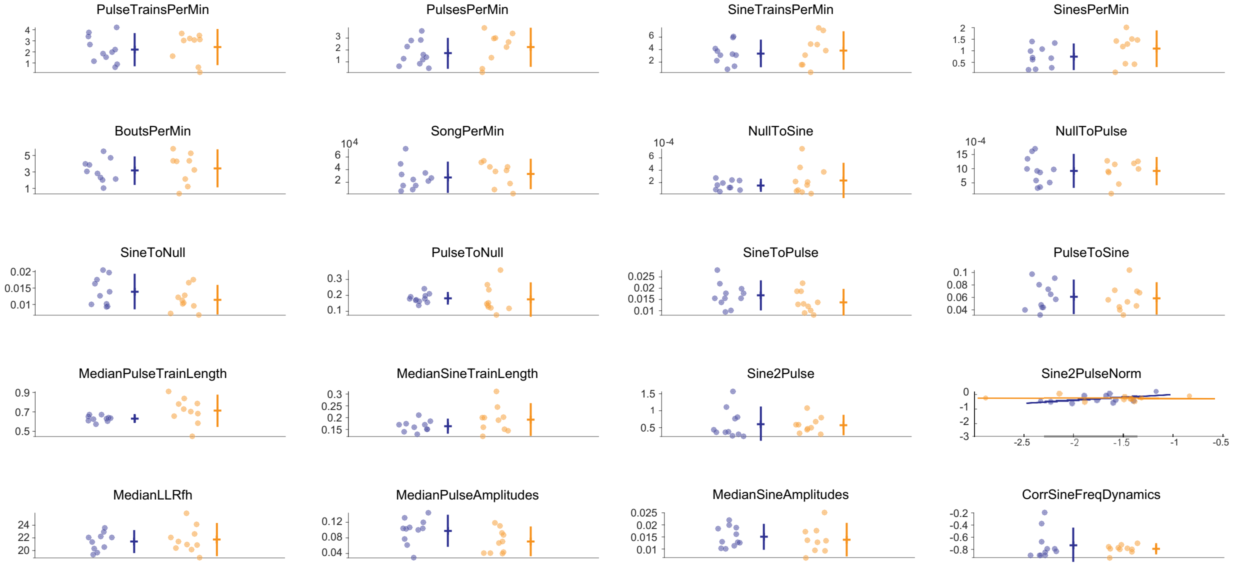
Extended Data Figure 3 | Targeted insertion of DsRed markers via CRISPR/Cas9-mediated HDR. a, **b**, Schematics of the targeted insertions of the markers 44.95$^{\text{DsRed}}$ (**a**) and 45.12$^{\text{DsRed}}$ (**b**). PAM, protospacer adjacent motif. **c**, **d**, PCR validation of 44.95$^{\text{DsRed}}$ (**c**) and 45.12$^{\text{DsRed}}$ (**d**) using the primer sets indicated in the panels above. A heterozygous fly was used for PCR and the negative control (NC) used a fly of the introgression line A2.3. The lower band (arrow) seen with primer set 2 therefore represents the wild-type allele. The PCR fragment containing the DsRed insertion is indicated by the white arrowhead. M, GeneRuler 1 kb DNA ladder.


a
gRNA
Donor ssDNA
Cas9 mRNA
lig4 DsiRNA
Co-injection
A2.3
Cross with *sim5*
Progeny screened by genotyping
EYFP+
or
EYFP-
Cross with *sim5*
mauW slo$^{\Delta1}$ allele
and
$slo^{\Delta1}$ detected by genotyping (EYFP linked to $slo^{\Delta1}$)
Check introgression breakpoints
Cross with *sim5*
Maintained by following EYFP through male lineage
sim5 slo$^{\Delta1}$ allele
and
$slo^{\Delta1}$ detected by genotyping
Cross with *sim5*
Maintained by genotyping each generation
b
Identity
Genomic DNA translation
TTTCGTATTCGAATATTGTGATTGCAGGGGATGTCGGGGTGTGATCAAAGCACTGTCGAATCATTGGCCGACGATC--CAACAGATTCACCATTCGACGCCGATGATTGTCTCAAGGTGCGCAAGTACTGGTGCTTTCTGCTGTCCAGCATCTT
M S G C D Q S T V E S L A D D P T D S P F D A D D C L K V R K Y W C F L L S S I
PAM
Protospacer
Donor ssDNA translation
ATTCGAATATTGTGATTGCAGGGGATGTCGGGGTGTGATCAAAGCACTGTCGAATCATTGGCCGACGATCGTTAACAGATTCACCATTCGACGCCGATGATTGTCTCAAGGTGCGCAAGTACTGGTGCTTTCTGCTGTCCAGC
M S G C D Q S T V E S L A D D R * Q I H H S T P M I V S R C A S T G A F C C P
sim5 slo$^{\Delta1}$ allele translation
TTTCGTATTCGAATATTGTGATTGCAGGGGATGTCGGGGTGTGATCAAAGCACTGTCGAATCATTGGCCGACGATCGTTAACAGATTCACCATTCGACGCCGATGATTGTCTCAAGGTGCGCAAGTACTGGTGCTTTCTGCTGTCCAGCATCTT
M S G C D Q S T V E S L A D D R * Q I H H S T P M I V S R C A S T G A F C C P A S
mauW slo$^{\Delta1}$ allele translation
TTTCGTATTCGAATATTGTGATTGCAGGGGATGTCGGGGTGTGATCAAAGCACTGTCGAATCATTGGCCGACGATCGTTAACAGATTCACCATTCGACGCCGATGATTGTCTCAAGGTGCGCAAGTACTGGTGCTTTCTGCTGTCCAGCATCTT
M S G C D Q S T V E S L A D D R * Q I H H S T P M I V S R C A S T G A F C C P A S

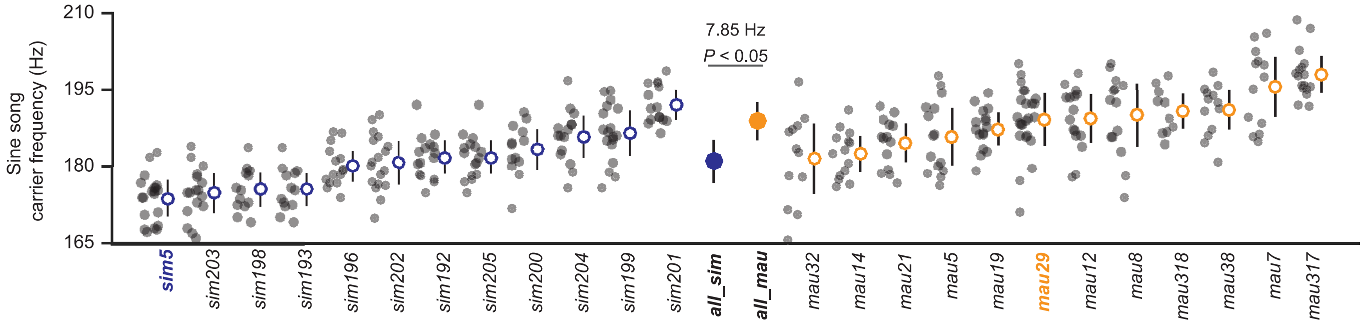
Extended Data Figure 4 | Generation and validation of the *slo*-null allele *slo*$^{\Delta1}$. a, Schematic of the generation of the *sim5* and *mauW slo*$^{\Delta1}$ alleles in the genetic background of the introgression line A2.3. Blue and orange bars represent *sim5* and *mauW* DNA, respectively. Green triangle denotes the EYFP marker. gRNA, guide RNA; ssDNA, single-stranded DNA; DsiRNA, Dicer-substrate short interfering RNA. **b**, Sequence verification of *slo*$^{\Delta1}$. Successful integration of ssDNA was confirmed by sequencing the cloned PCR products amplified from a heterozygous fly. Red asterisks indicate the introduced stop codon. PAM, protospacer adjacent motif.



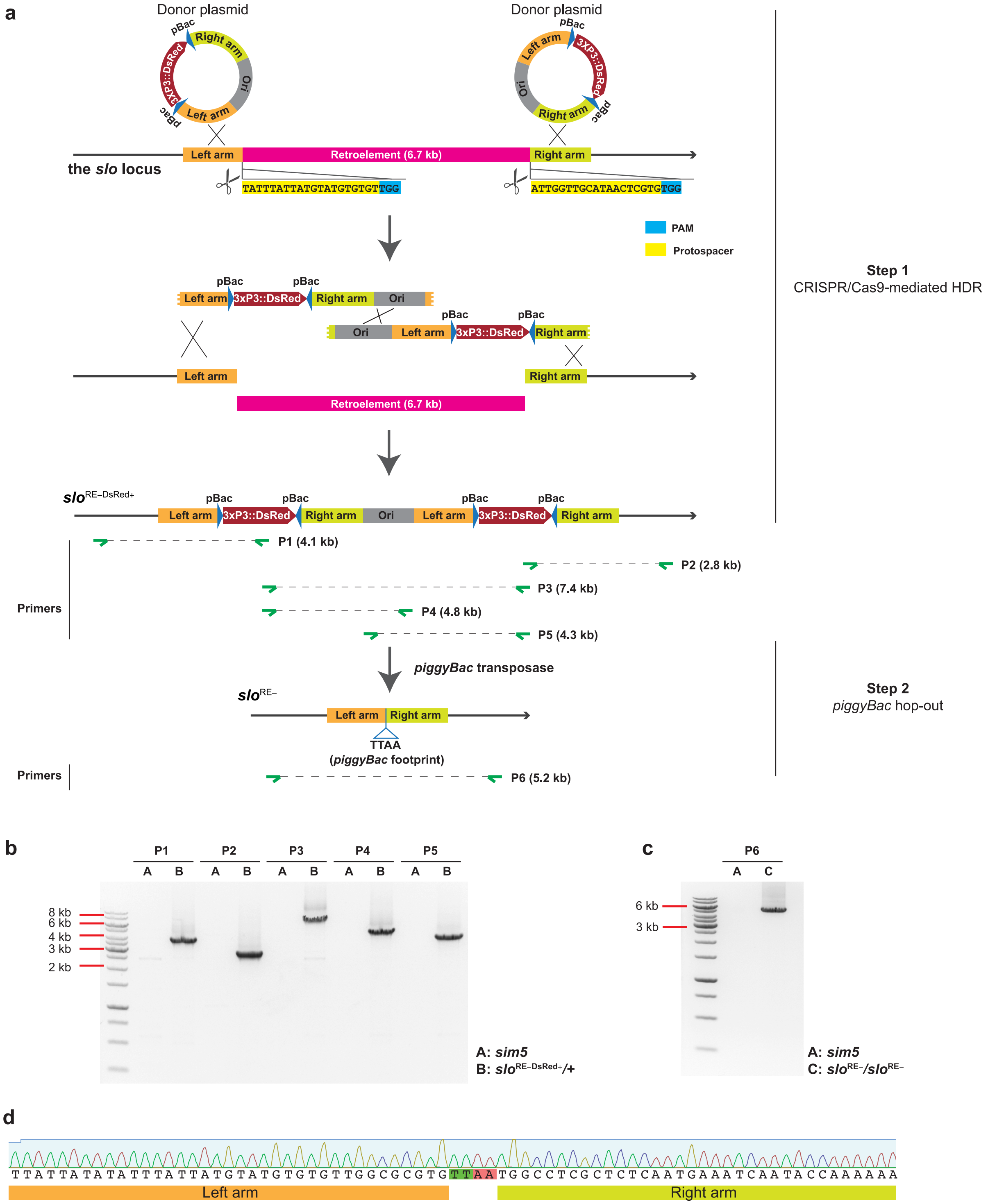
Extended Data Figure 5 | Expression pattern and song phenotype of *slo*. **a**, Schematic of the generation of the *slo*$^{\text{LexA::P65}}$ allele. T2A self-cleavage peptide sequences, LexA::P65, a 120 bp deletion, and a 3XP3::DsRed insertion were introduced into the first common exon shared by all *slo* transcripts in *D. melanogaster*. **b**, Schematic of the generation of the *slo*$^{\Delta 2}$ allele. Two stop codons, a 120 bp deletion, and a 3XP3::DsRed insertion were introduced into the same locus indicated in panel **a**. The same allele was generated in each of the following *D. simulans* strains: *sim5*, *sim202*, *sim203*, *sim205*. PAM, protospacer adjacent motif. **c**, **d**, Brain (**c**) and ventral nerve cord (**d**) of *D. melanogaster* LexAop-nls::GFP, *slo*$^{\text{LexA::P65}}$ males showed widespread expression of *slo*. **e**–**h**, Courtship song phenotype of *sim5 slo*$^{\Delta 2}$ males. The *slo*-null males produced very little song (**e**). SinesPerMin and PulsesPerMin measure the average amount of sine song and pulse song produced per minute, respectively. The sine and pulse events of the *slo*-null males were severely disrupted: sine song waveform (**f**); pulse train (**g**); pulse song waveform (**h**). **i**, *slo*$^{\Delta 2}$ heterozygotes sang sine song at 9.9 Hz lower frequency than wild-type. Data represent mean ± s.d. *P* values by one-way ANOVA.



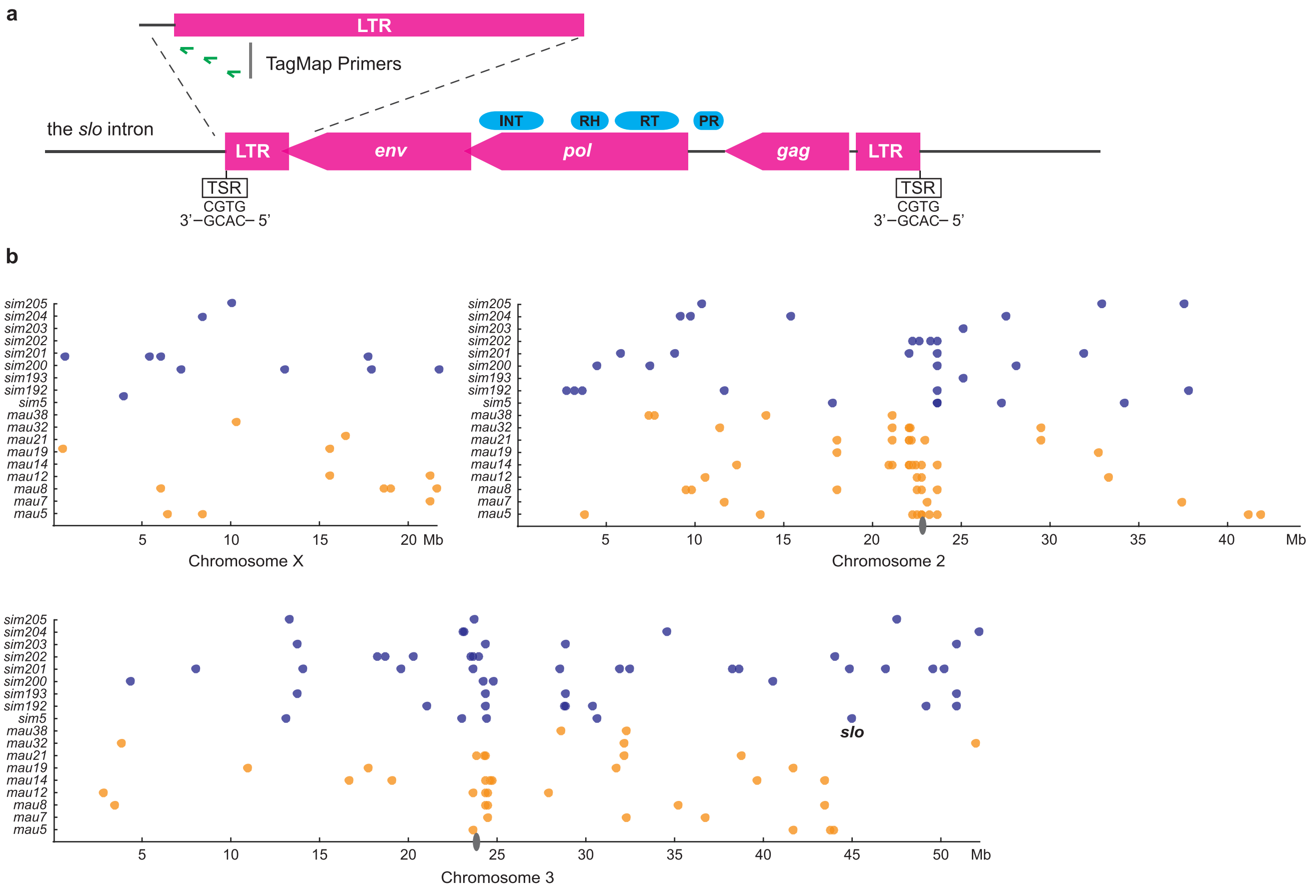
Extended Data Figure 6 | Additional song phenotypes of *slo* hemizygotes. Blue indicates the genotype *mauW*$^-$/*sim5*$^+$ and orange indicates *mauW*$^+$/*sim5*$^-$. None of these song phenotypes are significantly different by one-way ANOVA ($P > 0.05$). Data represent mean ± s.d. The definitions of each song phenotype are as follows: PulseTrainsPerMin, average number of pulse trains per minute; PulsesPerMin, average pulse song duration in seconds per minute; SineTrainsPerMin, average number of sine trains per minute; SinesPerMin, average sine song duration in seconds per minute; BoutsPerMin, average number of song bouts per minute; SongPerMin, average song duration per minute; NullToSine, transition probability from no song to sine song; NullToPulse, transition probability from no song to pulse song; SineToNull, transition probability from sine song to no song; PulseToNull, transition probability from pulse song to no song; SineToPulse, transition probability from sine song to pulse song within song bouts; PulseToSine, transition probability from pulse song to sine song within song bouts; MedianPulseTrainLength, median length of pulse trains; MedianSineTrainLength, median length of sine trains; Sine2Pulse, ratio of amount of sine song to amount of pulse song; Sine2PulseNorm, Sine2Pulse normalized by the amount of song; MedianLLRfh, median score of the log likelihood ratio of fit of pulse to pulse model versus white noise (measure of pulse shape); MedianPulseAmplitudes, median amplitude of pulses; MedianSineAmplitudes, median amplitude of sines; CorrSineFreqDynamics, slope of sine song carrier frequency within song bouts.



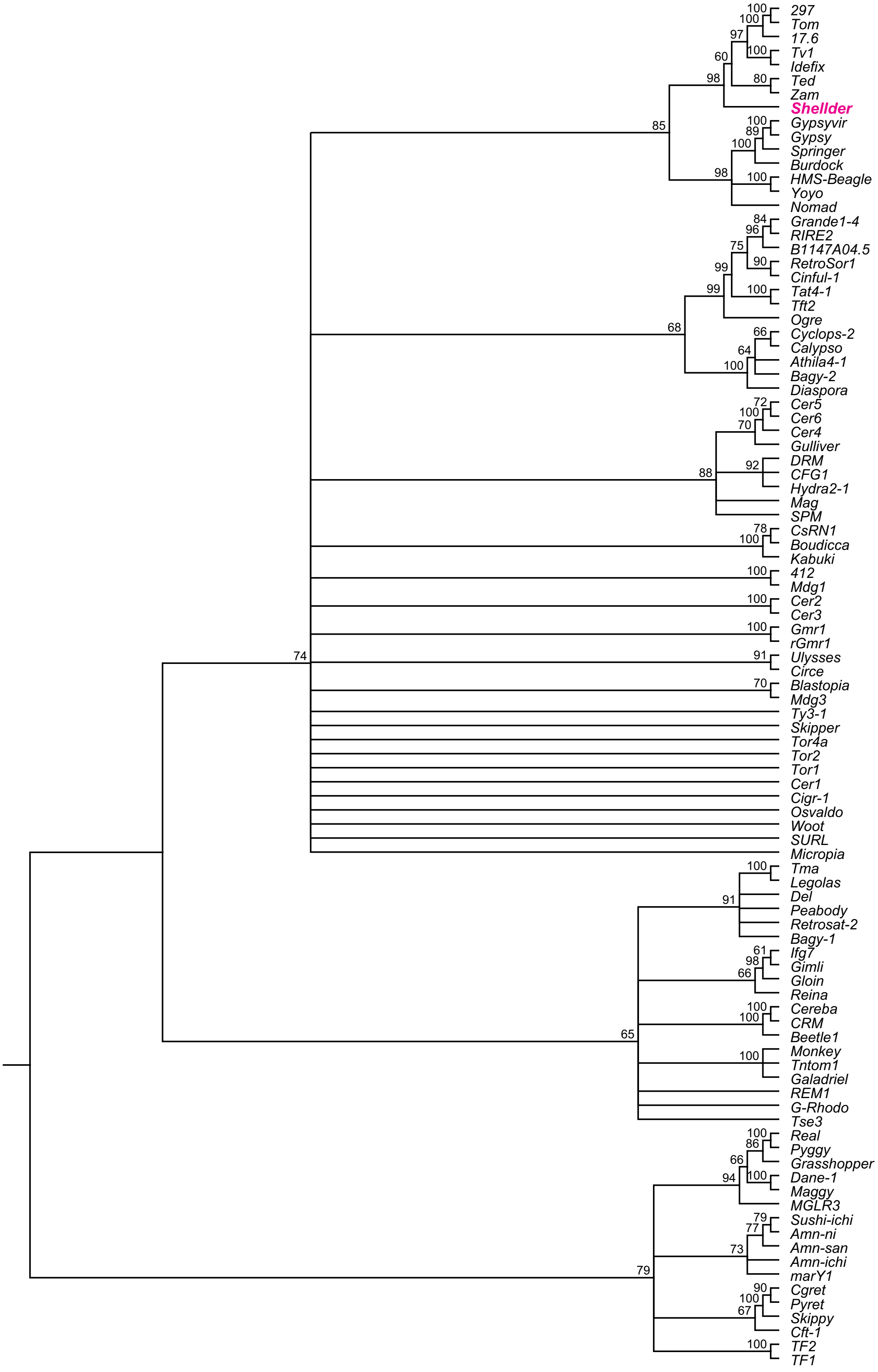
Extended Data Figure 7 | Sine song frequency phenotypes of 12 *D. simulans* wild-type isolates (blue) and 12 *D. mauritiana* wild-type isolates (orange). Open circles with lines represent mean ± s.d. Closed circles with lines represent mean ± s.d. of the means of all *D. simulans* (*all_sim*) and *D. mauritiana* strains (*all_mau*). *P* value by one-way ANOVA. All strains were recorded simultaneously through multiple recording sessions.



Extended Data Figure 8 | Targeted deletion of the retroelement insertion at the *slo* intron in *sim5*. a, Putative schematic of the targeted deletion of the retroelement insertion using a two-step strategy. First, the retroelement was replaced by a 3XP3::DsRed marker cassette flanked by *PiggyBac* transposon ends (labelled as pBac), using CRISPR/Cas9-mediated HDR. This generated the *slo*$^{\text{RE-DsRed+}}$ allele. Second, the 3XP3::DsRed marker was hopped out of the genome using *piggyBac* transposase to generate the *slo*$^{\text{RE}-}$ allele. In the first step, two independent HDR events appear to have occurred between the donor plasmids and the *slo* locus. All three *slo*$^{\text{RE-DsRed+}}$ alleles we generated support this rare recombination type (data not shown), which possibly reflects the difficulty of repairing a large deletion with a single donor plasmid. PAM, protospacer adjacent motif. **b**, **c**, PCR validation of the *slo*$^{\text{RE-DsRed+}}$ allele (**b**) and the *slo*$^{\text{RE}-}$ allele (**c**) using the primer sets indicated in panel **a**. M, GeneRuler 1 kb DNA ladder. **d**, Sequence verification of the *slo*$^{\text{RE}-}$ allele. The *piggyBac* footprint 'TTAA' is highlighted.



Extended Data Figure 9 | Identification of putative *Shellder* copies in *D. simulans* and *D. mauritiana* populations. **a**, Schematic of the *Shellder* insertion at the *slo* locus in *sim5*. *Shellder* contains three open reading frames (ORFs) resembling the *gag*, *pol*, and *env* genes of a retrovirus, flanked by 458 bp long terminal repeats (LTRs). Putative core protein domains are indicated: PR, protease; RT, reverse transcriptase; RH, RNase H; INT, integrase. In other retroviruses, the *pol* ORF often includes a 5′ protease domain. *Shellder* contains a conserved protease domain 5′ of the predicted *pol* start codon. It is possible that the *Shellder pol* ORF uses a non-ATG start codon. TSR, target site repeat. P1, P2, and P3 represent the three LTR-specific primers used for TagMap. **b**, Putative *Shellder* copies in *D. simulans* (blue) and *D. mauritiana* (orange) wild-type strains identified by TagMap. The *slo* locus insertion in *sim5* is indicated and is unique amongst these samples. *Shellder* insertions are enriched near centromeres (grey ovals), but can also be found in the euchromatic regions. Precise mapping locations are provided in Supplementary Table 5. This is probably an incomplete survey of *Shellder* copies, because TagMap may be biased towards detecting young and intact copies of transposable elements.



Extended Data Figure 10 | Phylogenetic position of *Shellder* in the Ty3/Gypsy family based on reverse transcriptase protein sequences. Bootstrap values (%) are indicated on the branches only when they exceed 60. Branch lengths are not drawn proportional to genetic distance.

doi:10.1038/nature19055

Operation of a homeostatic sleep switch

Diogo Pimentel[1]*, Jeffrey M. Donlea[1]*, Clifford B. Talbot[1], Seoho M. Song[1], Alexander J. F. Thurston[1] & Gero Miesenböck[1]

Sleep disconnects animals from the external world, at considerable risks and costs that must be offset by a vital benefit. Insight into this mysterious benefit will come from understanding sleep homeostasis: to monitor sleep need, an internal bookkeeper must track physiological changes that are linked to the core function of sleep[1]. In *Drosophila*, a crucial component of the machinery for sleep homeostasis is a cluster of neurons innervating the dorsal fan-shaped body (dFB) of the central complex[2,3]. Artificial activation of these cells induces sleep[2], whereas reductions in excitability cause insomnia[3,4]. dFB neurons in sleep-deprived flies tend to be electrically active, with high input resistances and long membrane time constants, while neurons in rested flies tend to be electrically silent[3]. Correlative evidence thus supports the simple view that homeostatic sleep control works by switching sleep-promoting neurons between active and quiescent states[3]. Here we demonstrate state switching by dFB neurons, identify dopamine as a neuromodulator that operates the switch, and delineate the switching mechanism. Arousing dopamine[4–8] caused transient hyperpolarization of dFB neurons within tens of milliseconds and lasting excitability suppression within minutes. Both effects were transduced by Dop1R2 receptors and mediated by potassium conductances. The switch to electrical silence involved the downregulation of voltage-gated A-type currents carried by Shaker and Shab, and the upregulation of voltage-independent leak currents through a two-pore-domain potassium channel that we term Sandman. Sandman is encoded by the *CG8713* gene and translocates to the plasma membrane in response to dopamine. dFB-restricted interference with the expression of Shaker or Sandman decreased or increased sleep, respectively, by slowing the repetitive discharge of dFB neurons in the ON state or blocking their entry into the OFF state. Biophysical changes in a small population of neurons are thus linked to the control of sleep–wake state.

We recorded from dFB neurons (which were marked by *R23E10-GAL4* or *R23E10-lexA*-driven green fluorescent protein (GFP) expression[3]) while head-fixed flies walked or rested on a spherical treadmill. Because inactivity is a necessary correlate but insufficient proof of sleep, we restricted our analysis to awakening, which we defined as a locomotor bout after ≥5 min of rest[9,10], during which the recorded dFB neuron had been persistently spiking. To deliver wake-promoting signals, we expressed the optogenetic actuator[5,11] CsChrimson under *TH-GAL4* control in the majority of dopaminergic neurons, including the PPL1 and PPM3 clusters[12], whose fan-shaped body (FB)-projecting members have been implicated in sleep control[4,8]. Illumination at 630 nm, sustained for 1.5 s to release a bolus of dopamine (Extended Data Fig. 1), effectively stimulated locomotion (32/38 trials; Fig. 1a, b). dFB neurons paused in successful (but not in unsuccessful) trials (Fig. 1a, b), and their membrane potentials dipped by 2–13 mV (7.50 ± 0.56 mV; mean ± standard error of the mean (s.e.m.)) below the baseline during tonic activity (Fig. 1a, c). When flies bearing an undriven *CsChrimson* transgene were photostimulated, neither physiological nor behavioural changes were apparent (Fig. 1d–f). The tight correlation between the suppression of dFB neuron spiking and the initiation of movement ($P < 0.0001$, Fisher's exact test) might, however, merely mirror a causal dopamine effect elsewhere, as *TH-GAL4* labels dopaminergic neurons throughout the brain[12]. Because localized dopamine applications to dFB neuron dendrites similarly caused awakening (see later), we consider this possibility remote.

Flies with enhanced dopaminergic transmission exhibit a short-sleeping phenotype that requires the presence of a D1-like receptor in dFB neurons[4,8], suggesting that dopamine acts directly on these cells. dFB-restricted RNA interference (RNAi) confirmed this notion and pinpointed Dop1R2 as the responsible receptor (Fig. 2a), a conclusion reinforced by analysis of the mutant *Dop1R2*^{MI08664} allele

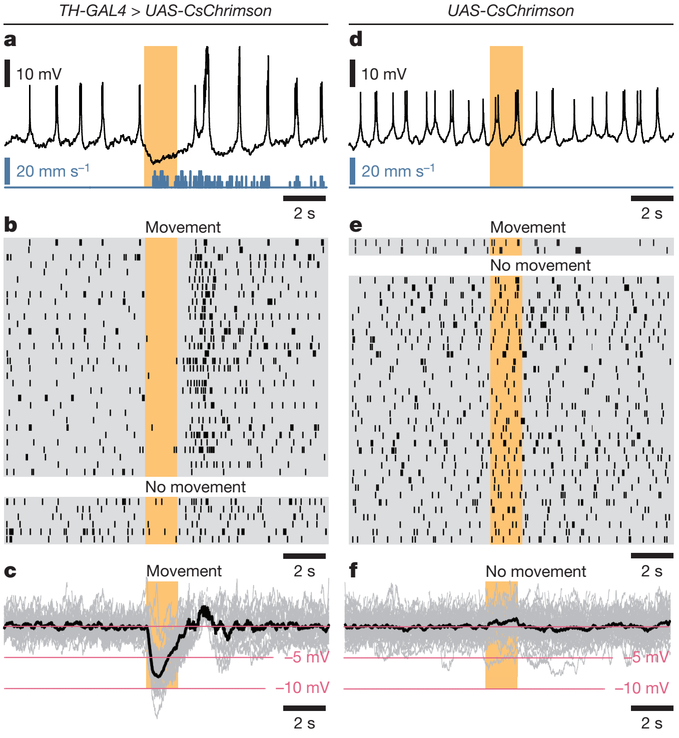


Figure 1 | Optogenetic stimulation of dopaminergic neurons silences dFB neurons and promotes awakening. a, Membrane potential (black) of a dFB neuron and simultaneously recorded movement (blue) of a fly expressing CsChrimson in dopaminergic neurons. **b**, Spike rasters of dFB neurons in 38 trials. Photostimulation elicited a behavioural response in 32 trials (top) and no response in 6 trials (bottom). **c**, Individual (grey) and average (black) membrane potentials during the 32 trials with a behavioural response. Spikes are blanked for clarity. **d**, Membrane potential (black) of a dFB neuron and simultaneously recorded movement (blue) of a fly lacking CsChrimson expression in dopaminergic neurons. **e**, Spike rasters of dFB neurons in 59 trials. Photostimulation elicited a behavioural response in 2 trials (top) and no response in 57 trials; of these, 36 were randomly selected for display (bottom). **f**, Individual (grey) and average (black) membrane potentials during the 57 trials without a behavioural response. Spikes are blanked for clarity.

[1]Centre for Neural Circuits and Behaviour, University of Oxford, Tinsley Building, Mansfield Road, Oxford OX1 3SR, UK.
*These authors contributed equally to this work.

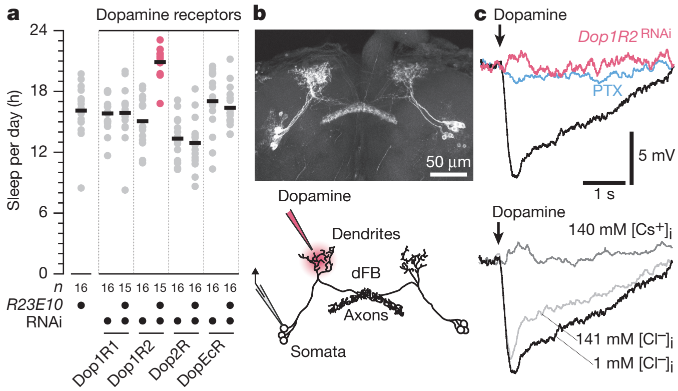


Figure 2 | Dopamine inhibits dFB neurons via Dop1R2 and the transient opening of a potassium conductance. a, Sleep in flies expressing *R23E10-GAL4*-driven RNAi targeting dopamine receptor transcripts and parental controls (circles, individual flies; horizontal lines, group means). One-way analysis of variance (ANOVA) detected a significant genotype effect ($P < 0.0001$); red indicates a significant difference from both parental controls in pairwise post-hoc comparisons. **b**, *R23E10-GAL4*-driven CD8::GFP expression in dFB neurons (top). Placement of pipettes for whole-cell recording and pharmacological stimulation (bottom). **c**, Membrane potentials of dFB neurons after a 250 ms pulse of dopamine, in control conditions of low intracellular chloride (1 mM; black, top and bottom); in cells expressing *R23E10-GAL4*-driven RNAi targeting *Dop1R2* (red, top); in the presence of 2 μg ml^{-1} intracellular PTX (blue, top); in elevated intracellular chloride (141 mM; light grey, bottom); and in intracellular caesium (140 mM; dark grey, bottom). Traces are averages of five dopamine applications.

(Extended Data Fig. 2a–c). Previous evidence that Dop1R1, a receptor not involved in regulating baseline sleep[8], confers responsiveness to dopamine when expressed in the dFB[4,8] indicates that either D1-like receptor can fulfil the role normally played by Dop1R2. Loss of Dop1R2 increased sleep during the day and the late hours of the night, by prolonging sleep bouts without affecting their frequency (Extended Data Fig. 2a, d, e). This sleep pattern is consistent with reduced sensitivity to a dopaminergic arousal signal.

To confirm the identity of the effective transmitter, avoid dopamine release outside the dFB, and reduce the transgene load for subsequent experiments, we replaced optogenetic manipulations of the dopaminergic system with pressure ejections of dopamine onto dFB neuron dendrites (Fig. 2b). Like optogenetically stimulated secretion, focal application of dopamine hyperpolarized the cells and suppressed their spiking (Fig. 2c and Extended Data Fig. 3a, b). The inhibitory responses could be blocked at several nodes of an intracellular signalling pathway that connects the activation of dopamine receptors to the opening of potassium conductances (Fig. 2c and Extended Data Fig. 3b): by RNAi-mediated knockdown of *Dop1R2*; by the inclusion in the patch pipette of pertussis toxin (PTX), which inactivates heterotrimeric G proteins of the $G_{i/o}$ family[13]; and by replacing intracellular potassium with caesium, which obstructs the pores of G-protein-coupled inward-rectifier channels[14]. Elevating the chloride reversal potential above resting potential left the polarity of the responses unchanged (Fig. 2c and Extended Data Fig. 3b), corroborating that potassium conductances mediate the bulk of dopaminergic inhibition.

Coupling of Dop1R2 to $G_{i/o}$, although documented in a heterologous system[15], represents a sufficiently unusual transduction mechanism for a predicted D1-like receptor to prompt us to verify its behavioural relevance. Like the loss of Dop1R2, temperature-inducible expression of PTX in dFB neurons increased overall sleep time by extending sleep bout length (Extended Data Fig. 2f, g).

While a single pulse of dopamine transiently hyperpolarized dFB neurons and inhibited their spiking, prolonged dopamine applications (50 ms pulses at 10 Hz, or 20 Hz optogenetic stimulation, both sustained for 2–10 min) switched the cells from electrical excitability (ON) to quiescence (OFF) (Fig. 3a–c and Extended Data Fig. 4a–c). The switching process required dopamine as well as Dop1R2 (Fig. 3b, c), but once the switch had been actuated the cells remained in the OFF state—and flies, awake (Fig. 3d)—without a steady supply of transmitter. Input resistances and membrane time constants dropped to 53.3 ± 1.8 and 24.0 ± 1.3% of their initial values (means ± s.e.m., $n = 15$ cells; Fig. 3b, c), and depolarizing currents no longer elicited action potentials (15 out of 15 cells) (Fig. 3a and Extended Data Fig. 4a). The biophysical properties of single dFB neurons, recorded in the same individual before and after operating the dopamine switch, varied as widely as those in sleep-deprived and rested flies[3].

Dopamine-induced changes in input resistance and membrane time constant occurred from similar baselines in all genotypes (Extended Data Fig. 5a, b) and followed single-exponential kinetics with time constants of 1.07–1.10 min (Fig. 3b, c). The speed of conversion points to post-translational modification and/or translocation of ion channels between intracellular pools and the plasma membrane as the underlying mechanism(s). In 7 out of 15 cases, we held recordings long enough to observe the spontaneous recommencement of spiking (Fig. 3a, d), which was accompanied by a rise to baseline of input resistance and membrane time constant, after 7–60 min of quiescence (mean ± s.e.m. = 25.86 ± 7.61 min). The temporary suspension of electrical output is thus part of the normal activity cycle of dFB neurons and not a dead end brought on by our experimental conditions.

dFB neurons in the ON state expressed two types of potassium current: voltage-dependent A-type[16] and voltage-independent non-A-type currents (Fig. 3e–g and Extended Data Fig. 6a–c). The current–voltage (I–V) relation of I_A resembled that of Shaker, the prototypical A-type channel[17,18]: no current flowed below −50 mV, the approximate voltage threshold of Shaker[17,18]; above −40 mV, peak currents increased steeply with voltage (Fig. 3e, f) and inactivated with a time constant[18] of 7.5 ± 2.1 ms (mean ± s.e.m., $n = 7$ cells; Extended Data Fig. 6c, d). Non-A-type currents showed weak outward rectification with a reversal potential of −80 mV (Fig. 3e, g), consistent with potassium as the permeant ion, and no inactivation (Extended Data Fig. 6b).

Switching the neurons OFF changed both types of potassium current. I_A diminished by one-third (Fig. 3e, f), whereas $I_{\text{non-A}}$ nearly quadrupled when quantified between resting potential and spike threshold (Fig. 3g). The weak rectification of $I_{\text{non-A}}$ in the ON state vanished in the OFF state, giving way to the linear I–V relationship of an ideal leak conductance (Fig. 3e, g). dFB neurons thus upregulate I_A in the sleep-promoting ON state (Fig. 3e, f). When dopamine switches the cells OFF, voltage-dependent currents are attenuated and leak currents augmented (Fig. 3e–g). This seesaw form of regulation should be sensitive to perturbations of the neurons' ion channel inventory: depletion of voltage-gated A-type (K_V) channels (which predominate in the ON state) should tip the cells towards the OFF state; conversely, loss of leak channels (which predominate in the OFF state) should favour the ON state. To test these predictions, we examined sleep in flies carrying *R23E10-GAL4*-driven RNAi transgenes for dFB-restricted interference with individual potassium channel transcripts.

RNAi-mediated knockdown of two of the five K_V channel types of *Drosophila*[19] (Shaker and Shab) reduced sleep relative to parental controls, while knockdown of the remaining three types had no effect (Fig. 4a). Biasing the potassium channel repertoire of dFB neurons against A-type conductances thus tilts the neurons' excitable state towards quiescence (Fig. 4b–f), causing insomnia (Fig. 4a), but leaves transient and sustained dopamine responses unaffected (Fig. 4e–g and Extended Data Fig. 3b). The seemingly counterintuitive conclusion that reducing a potassium current would decrease, not increase, action potential discharge is explained by a requirement for A-type channels in generating repetitive activity[16,20] of the kind displayed by dFB neurons during sleep (Fig. 1). Depleting Shaker from dFB neurons

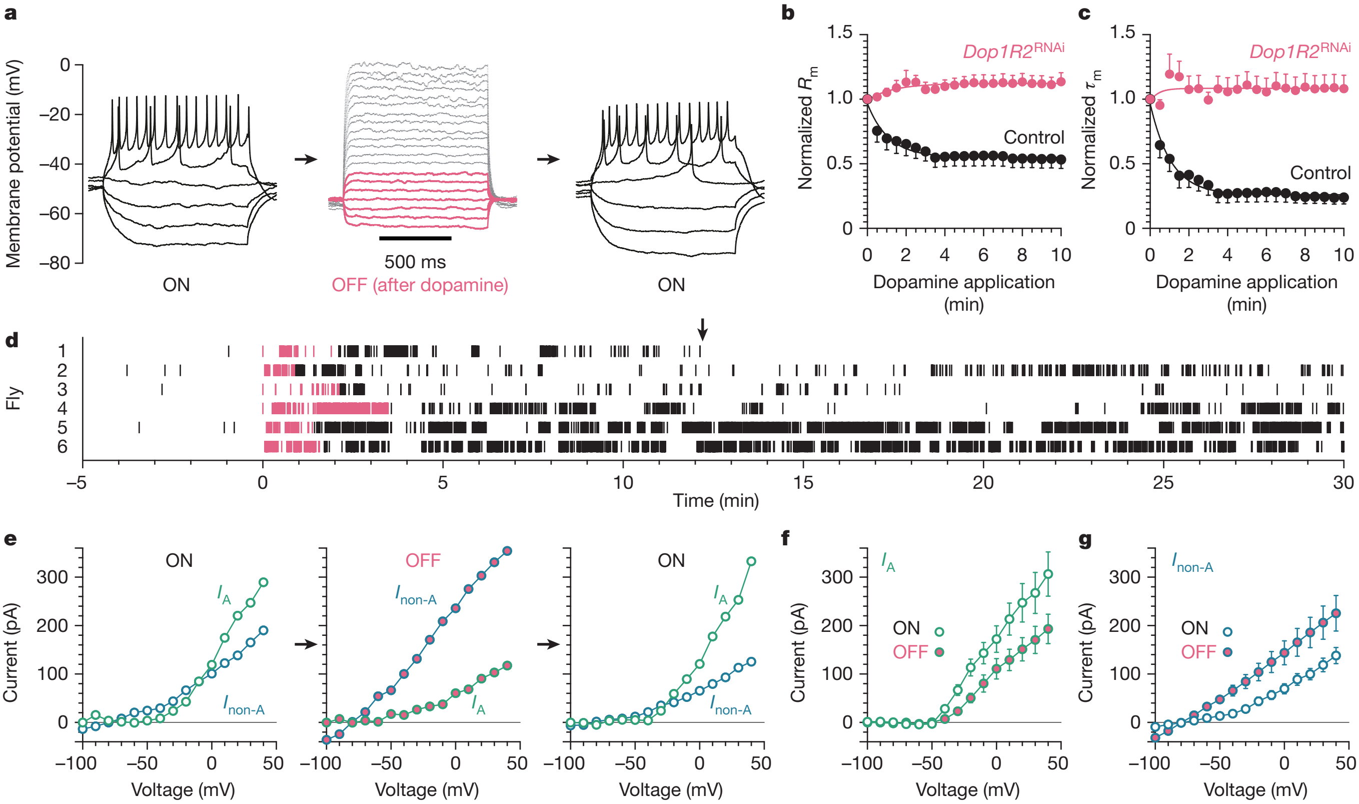


Figure 3 | Dopamine switches dFB neurons to quiescence via reciprocal modulation of two potassium conductances. a, A switching cycle in current clamp. Voltage responses to current steps were recorded in the same cell, before and after the application of dopamine. Red and grey traces in the OFF state (centre) indicate responses to current injections matching or exceeding those in the ON states, respectively. **b**, **c**, Time courses of changes in input resistance (R_m) and membrane time constant (τ_m) of dFB neurons during the application of dopamine, in controls (black, $n = 15$ cells) and cells expressing *R23E10-GAL4*-driven RNAi targeting *Dop1R2* (red, $n = 8$ cells). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected significant interactions between time and genotype ($P < 0.0001$ for R_m; $P < 0.0001$ for τ_m). **d**, Movement rasters of six flies before, during, and after bilateral applications of dopamine to dFB neuron dendrites. Vertical marks denote rotations of the treadmill (surface velocity $>4\,\text{mm s}^{-1}$, duration >50 ms). Red indicates the period of dopamine application, which started at 0 min (with the monitored dFB neuron in the ON state) and stopped when R_m fell to ~60% of its initial value. The arrow marks the spontaneous return to the ON state of the dFB neuron recorded in fly 1. Note the absence of movement thereafter. **e**, A switching cycle in voltage clamp. A-type (I_A, green) and non-A-type ($I_{non\text{-}A}$, blue) potassium currents evoked by voltage steps were recorded in the same cell, before and after the application of dopamine. **f**, Average ($n = 7$ cells) *I*–*V* relationships of I_A in the ON state (open symbols) and after dopamine-induced switching to the OFF state (red filled symbols). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected a significant interaction between voltage and neuronal state ($P < 0.0001$). **g**, Average ($n = 7$ cells) *I*–*V* relationships of $I_{non\text{-}A}$ in the ON state (open symbols) and after dopamine-induced switching to the OFF state (red filled symbols). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected a significant interaction between voltage and neuronal state ($P < 0.0001$).

shifted the interspike interval distribution towards longer values (Fig. 4d), as would be expected if K_V channels with slow inactivation kinetics replaced rapidly inactivating Shaker as the principal force opposing the generation of the next spike. These findings identify a potential mechanism for the short-sleeping phenotypes caused by mutations in *Shaker*[21], its β subunit *Hyperkinetic*[22], or its regulator *sleepless*[23] (Extended Data Fig. 7).

Leak conductances are typically formed by two-pore-domain potassium (K_{2P}) channels[24]. dFB-restricted RNAi of one member of the 11-strong family of *Drosophila* K_{2P} channels[19], encoded by the *CG8713* gene, increased sleep relative to parental controls; interference with the remaining 10 K_{2P} channels had no effect (Fig. 4a). Recordings from dFB neurons after knockdown of the *CG8713* gene product, which we term Sandman, revealed undiminished non-A-type currents in the ON state (Fig. 4c) and intact responses to a single pulse of dopamine (Fig. 4g and Extended Data Fig. 3b) but a defective OFF switch: during prolonged dopamine applications, $I_{non\text{-}A}$ failed to rise (Fig. 4c), input resistances and membrane time constants remained at their elevated levels (Fig. 4b, e, f and Extended Data Fig. 5), and the neurons continued to fire action potentials (7 out of 7 cells) (Fig. 4b). Blocking vesicle exocytosis in the recorded cell with botulinum neurotoxin C (BoNT/C)[25] similarly disabled the OFF switch (Fig. 4c, e, f). This, combined with the absence of detectable Sandman currents in the ON state (Fig. 4c), suggests that Sandman is internalized in electrically active cells and recycled to the plasma membrane when dopamine switches the neurons OFF.

Because dFB neurons lacking Sandman spike persistently even after prolonged dopamine exposure (Fig. 4b), voltage-gated sodium channels remain functional in the OFF state. The difficulty of driving control cells to action potential threshold in this state (Fig. 3a and Extended Data Fig. 4a) must therefore be due to a lengthening of electrotonic distance between sites of current injection and spike generation. This lengthening is an expected consequence of a current leak, which may uncouple the axonal spike generator from somatodendritic synaptic inputs or pacemaker currents when sleep need is low.

The two kinetically and mechanistically distinct actions of dopamine on dFB neurons—instant, but transient, hyperpolarization and a delayed, but lasting, switch in excitable state—ensure that transitions to vigilance can be both immediate and sustained, providing speedy alarm responses and stable homeostatic control. The key to stability lies in the switching behaviour of dFB neurons, which is

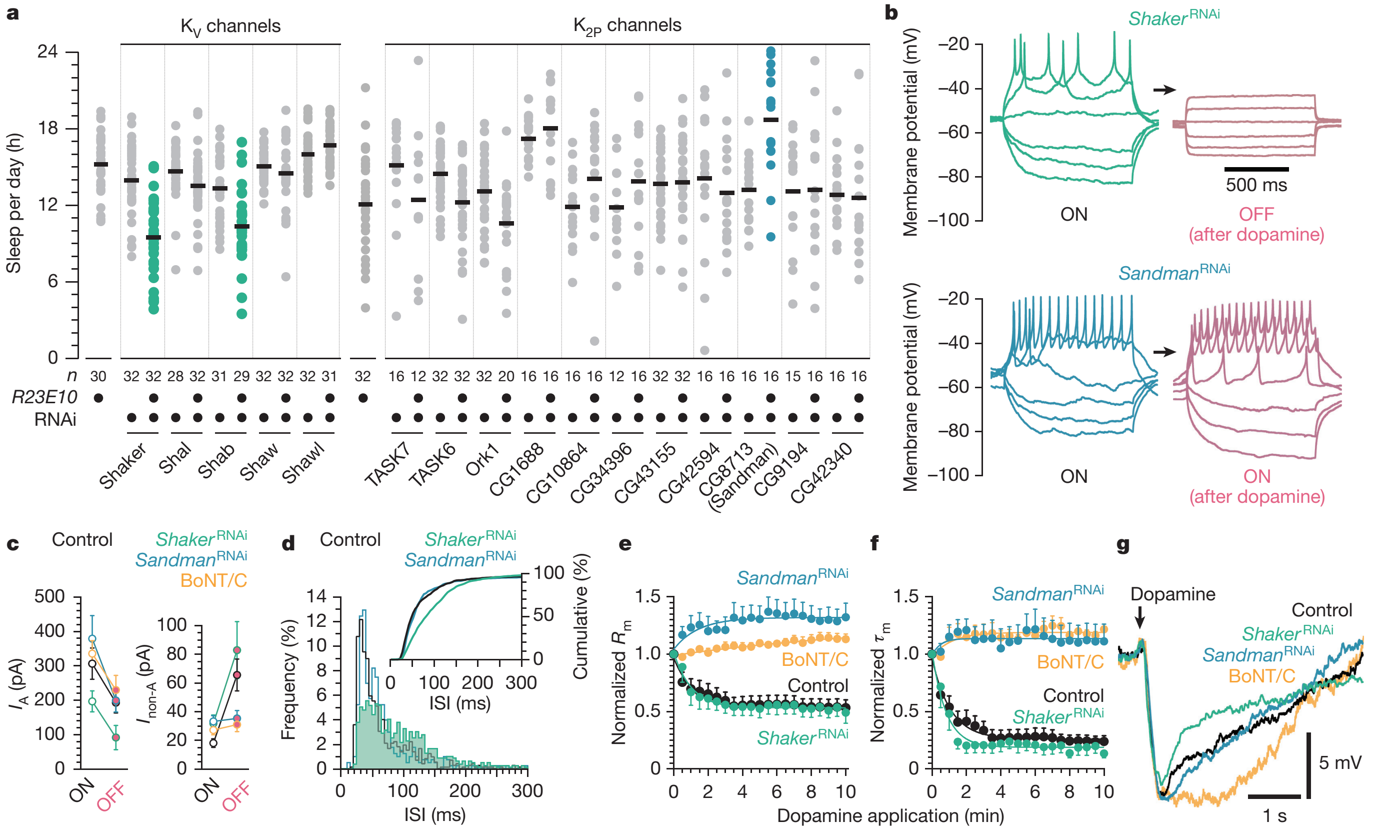


Figure 4 | The targets of antagonistic modulation by dopamine—Shaker and Sandman—have opposing effects on sleep. a, Sleep in flies expressing *R23E10-GAL4*-driven RNAi targeting K_V or K_{2P} channel transcripts and parental controls (circles, individual flies; horizontal lines, group means). One-way ANOVA detected significant genotype effects ($P < 0.0001$ for K_V channels; $P < 0.0001$ for K_{2P} channels); green and blue colours indicate significant differences from both parental controls in pairwise post-hoc comparisons. **b**, Voltage responses of two dFB neurons to current steps, before and after the application of dopamine. The neurons expressed *R23E10-GAL4*-driven RNAi targeting *Shaker* (green, top) or *Sandman* (blue, bottom). **c**, Amplitudes of I_A at 40 mV (left) and $I_{\text{non-A}}$ at −40 mV (right) in controls (black, $n = 7$ cells), neurons expressing *R23E10-GAL4*-driven RNAi targeting *Shaker* (green, $n = 7$ cells) or *Sandman* (blue, $n = 8$ cells), and in the presence of 1.5 μg ml^{-1} intracellular BoNT/C (orange, $n = 8$ cells), in the ON state (open symbols) and after dopamine-induced switching to the OFF state (red filled symbols). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected significant effects of experimental condition ($P = 0.0426$) and neuronal state ($P < 0.0001$) on I_A, and a significant interaction between experimental condition and neuronal state for $I_{\text{non-A}}$ ($P = 0.0018$). I_A was reduced in cells expressing *Shaker*$^{\text{RNAi}}$ relative to all other groups ($P = 0.0409$). $I_{\text{non-A}}$ differed between ON and OFF states in controls ($P = 0.0005$) and cells expressing *Shaker*$^{\text{RNAi}}$ ($P = 0.0003$), but not in cells expressing *Sandman*$^{\text{RNAi}}$ ($P = 0.9119$) or containing BoNT/C ($P = 0.9119$); $I_{\text{non-A}}$ in the ON state did not differ among groups ($P = 0.0782$). **d**, Frequency and cumulative frequency distributions (inset) of interspike intervals (ISIs) in controls (black) and neurons expressing *R23E10-GAL4*-driven RNAi targeting *Shaker* (green) or *Sandman* (blue). The interspike interval distribution of neurons expressing *Shaker*$^{\text{RNAi}}$ differed from that of the other groups ($P < 0.0001$ for both comparisons; Kolmogorov–Smirnov test). **e**, **f**, Time courses of changes in input resistance (R_m) and membrane time constant (τ_m) during the application of dopamine, in controls (black, $n = 15$ cells), neurons expressing *R23E10-GAL4*-driven RNAi targeting *Shaker* (green, $n = 6$ cells) or *Sandman* (blue, $n = 7$ cells), and in the presence of 1.5 μg ml^{-1} intracellular BoNT/C (orange, $n = 8$ cells). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected a significant interaction between time and experimental condition ($P < 0.0001$ for R_m; $P < 0.0001$ for τ_m). dFB neurons expressing *Sandman*$^{\text{RNAi}}$ or containing BoNT/C differed from controls ($P < 0.0001$ for all pairwise comparisons), but flies expressing *Shaker*$^{\text{RNAi}}$ did not ($P = 0.9993$ for R_m; $P = 0.8743$ for τ_m). **g**, Membrane potentials of dFB neurons after a 250 ms pulse of dopamine, in control flies (black), flies expressing *R23E10-GAL4*-driven RNAi targeting *Shaker* (green) or *Sandman* (blue), and in the presence of 1.5 μg ml^{-1} intracellular BoNT/C (orange). Traces are averages of five dopamine applications.

driven by dopaminergic input accumulated over time. Unlike bistable neurons[26–28], in which two activity regimes coexist for the same set of conductances, dFB neurons switch regimes only when their membrane current densities change. Our analysis of how dopamine effects such a change, from activity to silence, has uncovered elements familiar from other modulated systems[20,27–30]: simultaneous, antagonistic regulation of multiple conductances[20,29]; reduction of I_A (ref. 20); and modulation of leak currents[24]. We currently know little about the reverse transition, from silence to activity, except that mutating the Rho-GTPase-activating protein Crossveinless-c locks dFB neurons in the OFF state, resulting in severe insomnia and an inability to correct sleep deficits[3]. Discovering the signals and processes that switch sleep-promoting neurons back ON will hold important clues to the vital function of sleep.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 14 October 2015; accepted 24 June 2016.
Published online 3 August 2016.

1. Saper, C. B., Fuller, P. M., Pedersen, N. P., Lu, J. & Scammell, T. E. Sleep state switching. *Neuron* **68,** 1023–1042 (2010).
2. Donlea, J. M., Thimgan, M. S., Suzuki, Y., Gottschalk, L. & Shaw, P. J. Inducing sleep by remote control facilitates memory consolidation in *Drosophila*. *Science* **332,** 1571–1576 (2011).
3. Donlea, J. M., Pimentel, D. & Miesenböck, G. Neuronal machinery of sleep homeostasis in *Drosophila*. *Neuron* **81,** 860–872 (2014).
4. Liu, Q., Liu, S., Kodama, L., Driscoll, M. R. & Wu, M. N. Two dopaminergic neurons signal to the dorsal fan-shaped body to promote wakefulness in *Drosophila*. *Curr. Biol.* **22,** 2114–2123 (2012).

5. Lima, S. Q. & Miesenböck, G. Remote control of behavior through genetically targeted photostimulation of neurons. *Cell* **121,** 141–152 (2005).
6. Andretic, R., van Swinderen, B. & Greenspan, R. J. Dopaminergic modulation of arousal in *Drosophila*. *Curr. Biol.* **15,** 1165–1175 (2005).
7. Kume, K., Kume, S., Park, S. K., Hirsh, J. & Jackson, F. R. Dopamine is a regulator of arousal in the fruit fly. *J. Neurosci.* **25,** 7377–7384 (2005).
8. Ueno, T. *et al.* Identification of a dopamine pathway that regulates sleep and arousal in *Drosophila*. *Nature Neurosci.* **15,** 1516–1523 (2012).
9. Shaw, P. J., Cirelli, C., Greenspan, R. J. & Tononi, G. Correlates of sleep and waking in *Drosophila melanogaster*. *Science* **287,** 1834–1837 (2000).
10. Hendricks, J. C. *et al.* Rest in *Drosophila* is a sleep-like state. *Neuron* **25,** 129–138 (2000).
11. Zemelman, B. V., Lee, G. A., Ng, M. & Miesenböck, G. Selective photostimulation of genetically chARGed neurons. *Neuron* **33,** 15–22 (2002).
12. Claridge-Chang, A. *et al.* Writing memories with light-addressable reinforcement circuitry. *Cell* **139,** 405–415 (2009).
13. Bokoch, G. M., Katada, T., Northup, J. K., Ui, M. & Gilman, A. G. Purification and properties of the inhibitory guanine nucleotide-binding regulatory component of adenylate cyclase. *J. Biol. Chem.* **259,** 3560–3567 (1984).
14. Andrade, R. & Nicoll, R. A. Pharmacologically distinct actions of serotonin on single pyramidal neurones of the rat hippocampus recorded *in vitro*. *J. Physiol. (Lond.)* **394,** 99–124 (1987).
15. Reale, V., Hannan, F., Hall, L. M. & Evans, P. D. Agonist-specific coupling of a cloned *Drosophila melanogaster* D1-like dopamine receptor to multiple second messenger pathways by synthetic agonists. *J. Neurosci.* **17,** 6545–6553 (1997).
16. Connor, J. A. & Stevens, C. F. Prediction of repetitive firing behaviour from voltage clamp data on an isolated neurone soma. *J. Physiol. (Lond.)* **213,** 31–53 (1971).
17. Timpe, L. C. *et al.* Expression of functional potassium channels from Shaker cDNA in *Xenopus* oocytes. *Nature* **331,** 143–145 (1988).
18. Iverson, L. E., Tanouye, M. A., Lester, H. A., Davidson, N. & Rudy, B. A-type potassium channels expressed from Shaker locus cDNA. *Proc. Natl Acad. Sci. USA* **85,** 5723–5727 (1988).
19. Littleton, J. T. & Ganetzky, B. Ion channels and synaptic organization: analysis of the *Drosophila* genome. *Neuron* **26,** 35–43 (2000).
20. Harris-Warrick, R. M., Coniglio, L. M., Barazangi, N., Guckenheimer, J. & Gueron, S. Dopamine modulation of transient potassium current evokes phase shifts in a central pattern generator network. *J. Neurosci.* **15,** 342–358 (1995).
21. Cirelli, C. *et al.* Reduced sleep in *Drosophila* Shaker mutants. *Nature* **434,** 1087–1092 (2005).
22. Bushey, D., Huber, R., Tononi, G. & Cirelli, C. *Drosophila* hyperkinetic mutants have reduced sleep and impaired memory. *J. Neurosci.* **27,** 5384–5393 (2007).
23. Koh, K. *et al.* Identification of SLEEPLESS, a sleep-promoting factor. *Science* **321,** 372–376 (2008).
24. Enyedi, P. & Czirják, G. Molecular background of leak K^+ currents: two-pore domain potassium channels. *Physiol. Rev.* **90,** 559–605 (2010).
25. Schiavo, G., Matteoli, M. & Montecucco, C. Neurotoxins affecting neuroexocytosis. *Physiol. Rev.* **80,** 717–766 (2000).
26. Hounsgaard, J., Hultborn, H., Jespersen, B. & Kiehn, O. Bistability of α-motoneurones in the decerebrate cat and in the acute spinal cat after intravenous 5-hydroxytryptophan. *J. Physiol. (Lond.)* **405,** 345–367 (1988).
27. Marder, E., Abbott, L. F., Turrigiano, G. G., Liu, Z. & Golowasch, J. Memory from the dynamics of intrinsic membrane currents. *Proc. Natl Acad. Sci. USA* **93,** 13481–13486 (1996).
28. Marder, E. & Thirumalai, V. Cellular, synaptic and network effects of neuromodulation. *Neural Netw.* **15,** 479–493 (2002).
29. Baxter, D. A. & Byrne, J. H. Serotonergic modulation of two potassium currents in the pleural sensory neurons of Aplysia. *J. Neurophysiol.* **62,** 665–679 (1989).
30. Nicola, S. M., Surmeier, J. & Malenka, R. C. Dopaminergic modulation of neuronal excitability in the striatum and nucleus accumbens. *Annu. Rev. Neurosci.* **23,** 185–215 (2000).

Acknowledgements We thank S. Birman, R. Davis, B. Dickson, V. Jayaraman, L. Luo, G. Roman, G. Rubin, the Bloomington Stock Center, and the Vienna *Drosophila* Resource Center for flies. This work was supported by grants (to G.M.) from the Wellcome Trust, the Gatsby Charitable Foundation, the Oxford Martin School, and the National Institutes of Health. J.M.D. was the recipient of a postdoctoral fellowship from the Human Frontier Science Program; S.M.S. is a Commonwealth Scholar.

Author Contributions D.P., J.M.D. and G.M. designed the study and analysed the results. All electrophysiological recordings were done by D.P.; J.M.D. performed molecular manipulations and behavioural analyses with the help of S.M.S. and A.J.F.T. C.B.T. developed instrumentation. G.M. wrote the paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to G.M. (gero.miesenboeck@cncb.ox.ac.uk).

METHODS

***Drosophila* strains and culture.** Driver lines *R23E10-GAL4* or *R23E10-lexA*[31] and *TH-GAL4* (ref. 32) were used to target dFB neurons and dopaminergic neurons, respectively. Effector transgenes encoded fluorescent markers for visually guided patch-clamp recordings (*UAS-CD8::GFP*[33] and *lexAop-CD2::GFP*[34]); a temperature-inducible system[35] for the expression of pertussis toxin[36] (*UAS-PTX*; *tubP-GAL80*^{ts}); the optogenetic actuator CsChrimson[37]; and RNAi constructs[38], along with *UAS-Dcr2*, to interfere with the expression of the dopamine receptors Dop1R1 (107058KK), Dop1R2 (105324KK), DopR2 (11471GD) and DopEcR (103494KK); the K_V channels Shaker (104474KK), Shab (102218KK), Shal (103363KK), Shaw (110589KK) and Shawl (100980KK); the interacting partners of Shaker, Hyperkinetic (101402KK) and Sleepless (104533KK); and the K_{2P} channels Task7 (8565GD), Task6 (9073GD), Ork1 (104883KK), CG1688 (30270GD), CG10864 (8302GD), CG34396 (100436KK), CG43155 (101483KK), CG42594 (46415GD), CG8713 (47977GD), CG9194 (110628KK) and CG42340 (104521KK). Codes in parentheses identify transformants in the GD and KK libraries of the Vienna *Drosophila* Resource Center. The genotype of control flies in electrophysiological experiments was *w*^{1118}; *UAS-CD8::GFP*; *R23E10-GAL4*.

Fly stocks were grown on media of sucrose, yeast, molasses and agar under a 12 h light:12 h dark cycle at 25 °C unless they expressed GAL80^{ts}; in this case the experimental animals and all relevant controls were grown at 18 °C. Flies expressing CsChrimson were transferred to food supplemented with 2 mM all-*trans* retinal upon eclosion. All studies were performed on animals aged 3–10 days. Flies were routinely sleep-deprived[39] for >12 h before electrophysiological recordings to increase the likelihood of finding dFB neurons in the electrically active ON state after break-in.

Movement tracking, electrophysiology and optogenetics. Male and female flies with a dorsal cranial window were head-fixed to a custom mount, using thermoplastic wax with a melting point of 52 °C (Agar Scientific), and placed on a spherical treadmill[40,41]. The treadmill consisted of an air-supported trackball made of extruded styrofoam (13 mm diameter; 50 mg) in a 14 mm tube. An image of a small region of the ball's surface under 640 nm LED illumination was relayed onto the sensor of an optical mouse (Logitec M-U0017). The sensor was interfaced with a microcontroller board (Arduino Due) based on the Atmel SAM3X CPU and read out in real time using the onboard D/A converter. The resolution of the readout corresponds to 4 mm s^{-1} increments in the tangential speed of the trackball.

The brain was continuously superfused with extracellular solution equilibrated with 95% O_2–5% CO_2 and containing 103 mM NaCl, 3 mM KCl, 5 mM TES, 8 mM trehalose, 10 mM glucose, 7 mM sucrose, 26 mM $NaHCO_3$, 1 mM NaH_2PO_4, 1.5 mM $CaCl_2$, 4 mM $MgCl_2$, pH 7.3. Somata of GFP-labelled dFB neurons were visually targeted with borosilicate glass electrodes (7–13 MΩ). The internal solution contained 140 mM potassium aspartate, 10 mM HEPES, 1 mM KCl, 4 mM MgATP, 0.5 mM Na_3GTP, 1 mM EGTA, pH 7.3. Where indicated, 140 mM potassium aspartate was replaced with 140 mM KCl or 140 mM caesium aspartate, or the internal solution was supplemented with 2 μg ml^{-1} pertussis toxin (Tocris) or 1.5 μg ml^{-1} botulinum neurotoxin C light chain (List Biological Laboratories). Neurons were dialysed with toxin-containing internal solutions for 10–15 min before measurements.

Signals were acquired with a Multiclamp 700B amplifier (Molecular Devices), filtered at 6–10 kHz, and digitized at 10–20 kHz using an ITC-18 data acquisition board (InstruTECH) controlled by the Nclamp/Neuromatic package. Data were analysed using Neuromatic software (http://www.neuromatic.thinkrandom.com) and custom procedures in Igor Pro (Wavemetrics) and MATLAB (The MathWorks).

Voltage-clamp experiments were performed in the presence of 1 μM tetrodotoxin (Tocris) and 200 nM cadmium to block sodium and calcium channels, respectively. Neurons were taken in 10 mV increments from holding potentials of −110 or −30 mV to test potentials between −100 and 40 mV (Extended Data Fig. 6a, b). When the cells were held at −110 mV, depolarization steps (1 s duration) elicited the full complement of potassium currents; when the cells were held at −30 mV, voltage-gated channels inactivated and the evoked potassium currents lacked the I_A (A-type or fast outward) component[42]. The non-A-type component was quantified at the steady state (end) of the current response. Digital subtraction of the non-A-type component from the full complement of potassium currents (that is, the currents evoked from a hyperpolarized holding potential of −110 mV) gave an estimate of I_A (Extended Data Fig. 6c), which was taken to be the difference between the peak current and any residual steady state current in the difference trace.

Interspike intervals were determined from voltage responses to a standard series of depolarizing current steps (5 pA increments from 0 to 150 pA, 1 s duration). Spikes were detected by finding minima in the second derivative of the membrane potential trace. Interspike intervals at all levels of injected current were pooled for the calculation of frequency distributions.

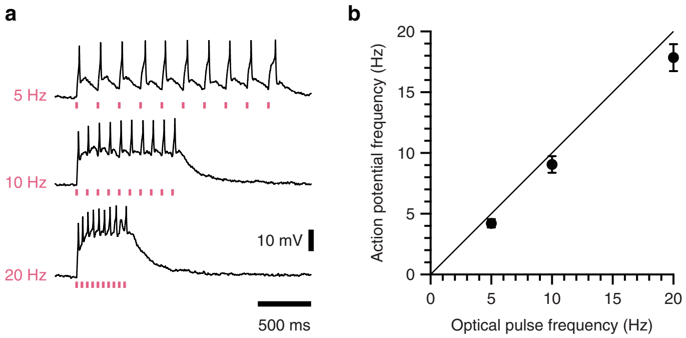
For photostimulation of CsChrimson-expressing neurons[37], a 630 nm LED (Multicomp OSW-4388) was focused onto the head of the fly with a 60 mm lens (Thorlabs) and controlled by a TTL-triggered dimmable constant current LED driver (Recom RCD-24-0.70/W/X3). Optical power at the sample was ~28 mW cm^{-2}. To induce state switching, light was delivered in a pulsatile fashion in 5 s cycles. The first 3.5 s of each cycle consisted of 20 Hz trains of 3 ms optical pulses. Illumination was paused during the remaining 1.5 s of each cycle, and a hyperpolarizing current pulse (−10 pA; 1 s) was applied to determine the membrane resistance and time constant.

For pharmacological applications of dopamine, patch pipettes were filled with 10 mM dopamine in extracellular solution and positioned in the centre of the GFP-labelled dendritic tuft of dFB neurons. To elicit transient dopamine responses, pressure (68 kPa) was applied in 250 ms pulses (Picospritzer III), resulting in the ejection of ~40 pl of solution. To induce state switching, dopamine was delivered in a pulsatile fashion in 5 s cycles. The first 3.5 s of each cycle consisted of 10 Hz trains of 50 ms pressure pulses. Dopamine delivery was paused during the remaining 1.5 s of each cycle, and a hyperpolarizing current pulse (−10 pA; 1 s) was applied to determine the membrane resistance and time constant.

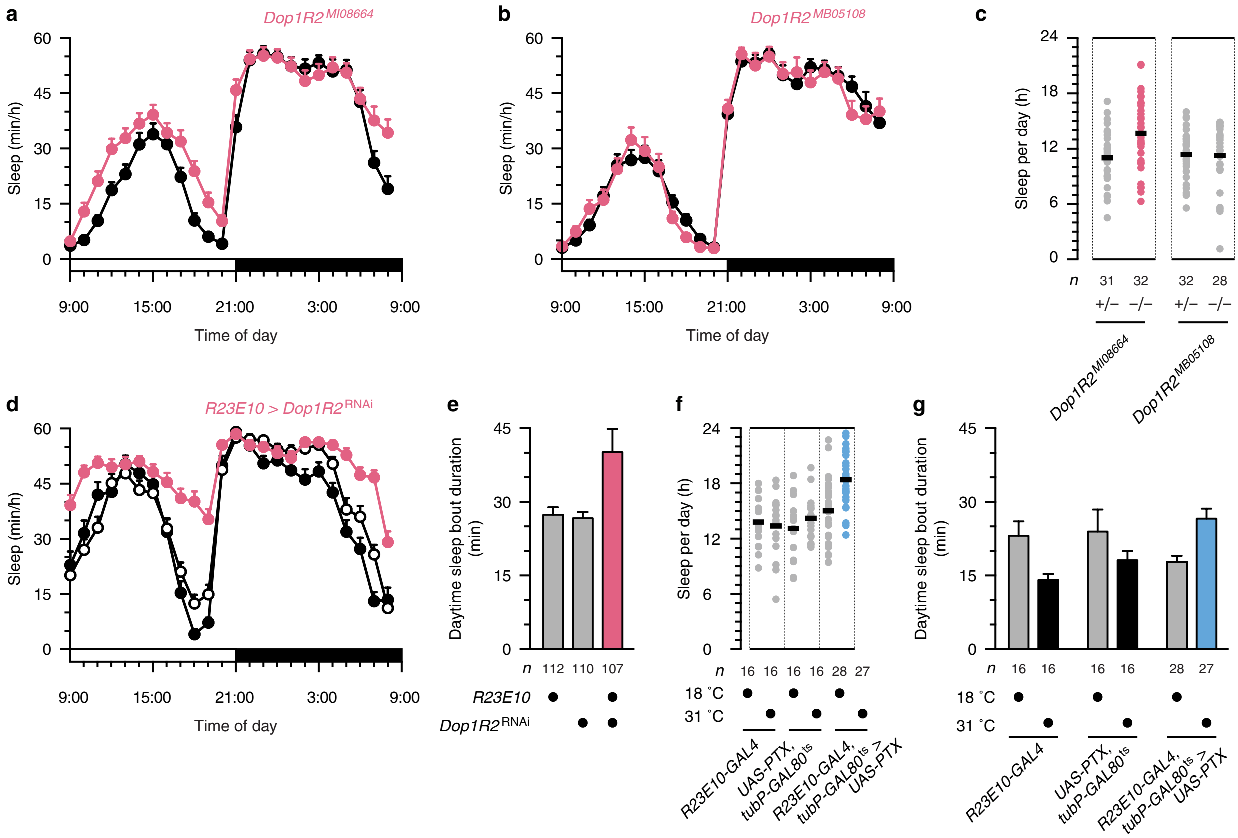
Sleep measurements. Female flies were individually inserted into 65 mm glass tubes, loaded into the Trikinetics *Drosophila* Activity Monitor system, and housed under 12 h light:12 h dark conditions. Periods of inactivity (no beam breaks) lasting at least 5 min were classified as sleep[9,10]. Immobile flies (<2 beam breaks per 24 h) were excluded from analysis. Group sizes for sleep measurements (typically $n = 16$ flies; in some cases multiples of 16) reflect the capacity of the Trikinetics *Drosophila* Activity Monitors, which were designed to accommodate 16 experimental flies along with 16 controls.

Statistics. Data were analysed in Prism 6 (GraphPad). Group means were compared by one-way or two-way ANOVA, using repeated measures designs where appropriate, followed by planned pairwise post-hoc analyses using Holm–Šídák's multiple comparisons test. Where the assumptions of normality or sphericity were violated (as indicated by Shapiro–Wilk and Brown–Forsythe tests, respectively), group means were compared by two-sided Mann–Whitney or Kruskal–Wallis tests, the latter followed by Dunn's multiple comparisons test. Contingencies between the suppression of dFB neuron activity and awakening were analysed by Fisher's exact test. Interspike interval distributions were evaluated by Kolmogorov–Smirnov test, using the Bonferroni correction to adjust the level of statistical significance. No statistical methods were used to predetermine sample sizes. The experiments were not randomized, and the investigators were not blinded to allocation during experiments and outcome assessment.

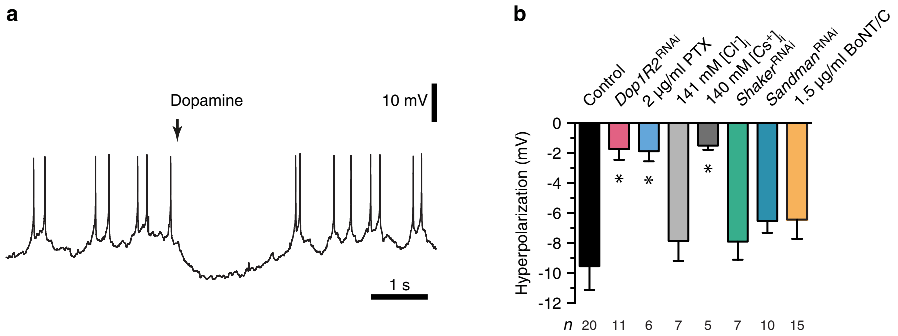
31. Jenett, A. *et al.* A GAL4-driver line resource for *Drosophila* neurobiology. *Cell Reports* **2,** 991–1001 (2012).
32. Friggi-Grelin, F. *et al.* Targeted gene expression in *Drosophila* dopaminergic cells using regulatory sequences from tyrosine hydroxylase. *J. Neurobiol.* **54,** 618–627 (2003).
33. Lee, T. & Luo, L. Mosaic analysis with a repressible cell marker for studies of gene function in neuronal morphogenesis. *Neuron* **22,** 451–461 (1999).
34. Pfeiffer, B. D. *et al.* Refinement of tools for targeted gene expression in *Drosophila*. *Genetics* **186,** 735–755 (2010).
35. McGuire, S. E., Le, P. T., Osborn, A. J., Matsumoto, K. & Davis, R. L. Spatiotemporal rescue of memory dysfunction in *Drosophila*. *Science* **302,** 1765–1768 (2003).
36. Ferris, J., Ge, H., Liu, L. & Roman, G. G(o) signaling is required for *Drosophila* associative learning. *Nature Neurosci.* **9,** 1036–1040 (2006).
37. Klapoetke, N. C. *et al.* Independent optical excitation of distinct neural populations. *Nature Methods* **11,** 338–346 (2014).
38. Dietzl, G. *et al.* A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448,** 151–156 (2007).
39. Shaw, P. J., Tononi, G., Greenspan, R. J. & Robinson, D. F. Stress response genes protect against lethal effects of sleep deprivation in *Drosophila*. *Nature* **417,** 287–291 (2002).
40. Buchner, E. Elementary movement detectors in an insect visual-system. *Biol. Cybern.* **24,** 85–101 (1976).
41. Seelig, J. D. *et al.* Two-photon calcium imaging from head-fixed *Drosophila* during optomotor walking behavior. *Nature Methods* **7,** 535–540 (2010).
42. Connor, J. A. & Stevens, C. F. Voltage clamp studies of a transient outward membrane current in gastropod neural somata. *J. Physiol. (Lond.)* **213,** 21–30 (1971).



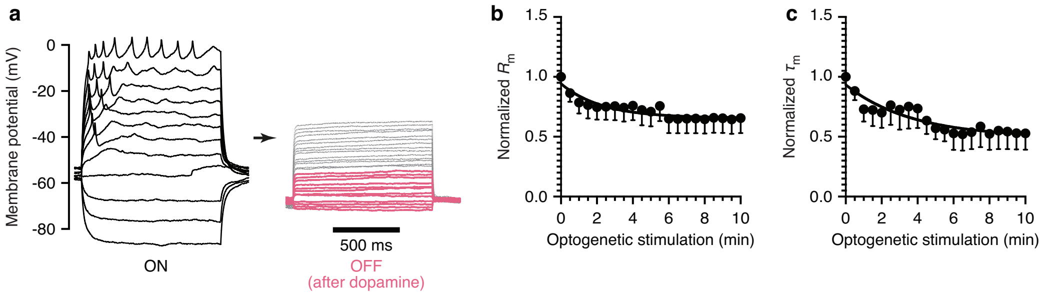
Extended Data Figure 1 | Optogenetic stimulation of dopaminergic neurons. Dopaminergic neurons expressing CsChrimson under *TH-GAL4* control were driven with 3 ms pulses of 630 nm light at the indicated frequencies. Optical power at the sample was ~28 mW cm^{-2}. **a**, Examples of voltage responses to optical pulse trains. **b**, The ratio of light-evoked action potentials to optical pulses was close to 1 at driving frequencies between 5 and 20 Hz ($n = 36$ trials on 6 cells). Data are means ± s.e.m.



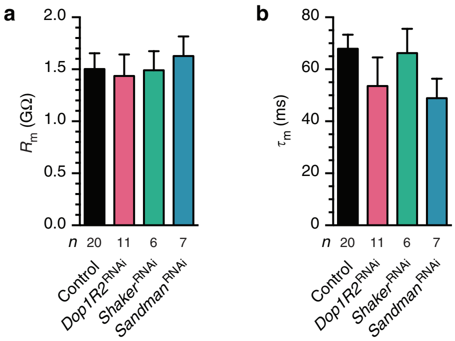
Extended Data Figure 2 | Changes in sleep after interference with Dop1R2 signalling are consistent with diminished sensitivity to arousing dopamine. **a**, Sleep during a 24 h day in homozygous carriers of the *Dop1R2*^{MI08664} allele (red, $n = 32$ flies) and heterozygous controls (black, $n = 31$ flies). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected a significant interaction between time of day and genotype ($P < 0.0001$). **b**, Sleep during a 24 h day in homozygous carriers of the *Dop1R2*^{MB05108} allele (red, $n = 28$ flies) and heterozygous controls (black, $n = 32$ flies). Data are means ± s.e.m. Two-way repeated-measures ANOVA failed to detect a significant interaction between time of day and genotype ($P = 0.4736$). **c**, Sleep in homozygous and heterozygous carriers of the *Dop1R2*^{MI08664} or *Dop1R2*^{MB05108} alleles (circles, individual flies; horizontal lines, group means). Mann–Whitney tests detected a significant effect of the *Dop1R2*^{MI08664} allele ($P = 0.0219$, red), but not of the *Dop1R2*^{MB05108} allele ($P = 0.6750$). The *Dop1R2*^{MB05108} allele contains a transposon insertion in a non-coding region of the *Dop1R2* gene, which reduces mRNA levels in homozygous carriers by only 14% (ref. 4), thus explaining the lack of a phenotype. The inability of *Dop1R2*^{MB05108} to suppress the short-sleeping phenotype of flies with enhanced dopaminergic transmission[4] therefore does not argue against a role of Dop1R2 in the dFB. **d**, Sleep during a 24 h day in flies expressing *R23E10-GAL4*-driven RNAi targeting *Dop1R2* (red, $n = 48$ flies) and parental controls (open symbols: *R23E10-GAL4*, $n = 48$ flies; filled symbols: undriven *UAS-Dop1R2*^{RNAi}, $n = 32$ flies). Data are means ± s.e.m. Two-way repeated-measures ANOVA detected a significant interaction between time of day and genotype ($P < 0.0001$). **e**, Average length of daytime sleep bouts in flies expressing *R23E10-GAL4*-driven RNAi targeting *Dop1R2* and parental controls. Data are means ± s.e.m. One-way ANOVA detected a significant genotype effect ($P = 0.0015$); red indicates a significant difference from both parental controls in pairwise post-hoc comparisons. **f**, Sleep in flies with temperature-inducible *R23E10-GAL4*-driven expression of PTX and parental controls (circles, individual flies; horizontal lines, group means). Two-way ANOVA detected a significant interaction between genotype and temperature ($P = 0.0143$); blue indicates a significant difference between inducing and non-inducing temperatures in pairwise post-hoc comparisons. **g**, Average length of daytime sleep bouts in flies with temperature-inducible *R23E10-GAL4*-driven expression of PTX and parental controls. Data are means ± s.e.m. Two-way ANOVA detected a significant interaction between genotype and temperature ($P = 0.0002$); blue indicates a significant increase upon switching from non-inducing to inducing temperatures in pairwise post-hoc comparisons.



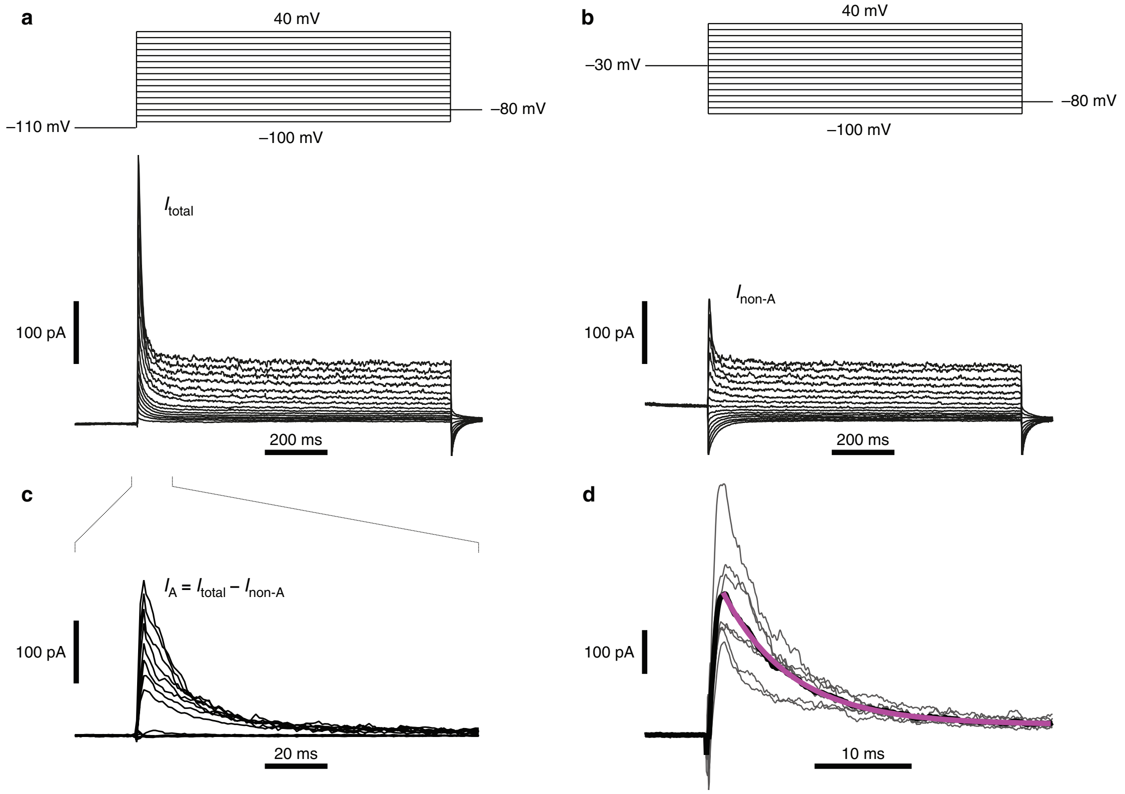
Extended Data Figure 3 | Dopamine hyperpolarizes dFB neurons and inhibits their spiking. a, Membrane potential of a dFB neuron during a 250 ms pulse of dopamine. **b**, Average amplitude of hyperpolarization evoked by dopamine in the indicated numbers of cells. Data are means ± s.e.m. Kruskal–Wallis test detected a significant difference between groups ($P < 0.0001$); asterisks indicate significant differences from controls in pairwise post-hoc comparisons.



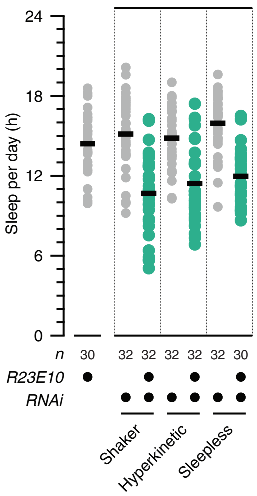
Extended Data Figure 4 | Optogenetic stimulation of dopaminergic neurons switches dFB neurons to quiescence. Flies expressing CsChrimson under *TH-GAL4* control in dopaminergic neurons were photostimulated with 3 ms pulses of 630 nm light at 20 Hz. **a**, Voltage responses to current steps were recorded in the same cell, before and after optogenetic stimulation of dopaminergic neurons (black and red traces). Red and grey traces in the OFF state (right) indicate current injections matching or exceeding those in the ON state, respectively (left). **b**, **c**, Time courses of changes in input resistance (R_{m}) and membrane time constant (τ_{m}) of dFB neurons during optogenetic stimulation of dopaminergic neurons ($n = 7$ cells). Data are means ± s.e.m. One-way repeated-measures ANOVA detected significant effects of time ($P = 0.0135$ for R_{m}; $P = 0.0222$ for τ_{m}).



Extended Data Figure 5 | Membrane properties of dFB neurons in the ON state. a, Input resistances (R_m) of the indicated numbers of cells. Data are means ± s.e.m. Kruskal–Wallis test failed to detect a significant difference between groups ($P = 0.8997$). **b**, Membrane time constants (τ_m) of the indicated numbers of cells. Data are means ± s.e.m. Kruskal–Wallis test failed to detect a significant difference between groups ($P = 0.1682$).



Extended Data Figure 6 | Measurements of potassium currents in voltage clamp. a, Voltage steps from a holding potential of −110 mV (top) elicited the full complement of potassium currents expressed by a dFB neuron (I_{total}, bottom). **b**, Stepping the same neuron from a holding potential of −30 mV (top) elicited potassium currents lacking the A-type component ($I_{non\text{-}A}$, bottom). **c**, Digital subtraction of $I_{non\text{-}A}$ (**b**, bottom) from I_{total} (**a**, bottom) yielded an estimate of I_A. Note the expanded timescale. **d**, Individual (grey) and average (black) A-type currents of seven dFB neurons, evoked by step depolarization to 40 mV. The magenta line represents a single-exponential fit to the average.



Extended Data Figure 7 | Loss of Shaker and its interacting partners, Hyperkinetic and Sleepless, from dFB neurons has similar effects on sleep. Sleep in flies expressing *R23E10-GAL4*-driven RNAi targeting *Shaker*, *Hyperkinetic* or *sleepless* and parental controls (circles, individual flies; horizontal lines, group means). One-way ANOVA detected a significant genotype effect ($P < 0.0001$); green indicates significant differences from both parental controls in pairwise post-hoc comparisons.

LETTER

doi:10.1038/nature19067

A human neurodevelopmental model for Williams syndrome

Thanathom Chailangkarn[1,2,3,4]*, Cleber A. Trujillo[1,2,3]*, Beatriz C. Freitas[1,2,3], Branka Hrvoj-Mihic[5], Roberto H. Herai[1,2,3,6], Diana X. Yu[7], Timothy T. Brown[8,9,10], Maria C. Marchetto[7], Cedric Bardy[7,11], Lauren McHenry[7], Lisa Stefanacci[1,2,3,5]‡, Anna Järvinen[12], Yvonne M. Searcy[12], Michelle DeWitt[12], Wenny Wong[12], Philip Lai[12], M. Colin Ard[9], Kari L. Hanson[5], Sarah Romero[1,2,3], Bob Jacobs[13], Anders M. Dale[8,14,15], Li Dai[16,17], Julie R. Korenberg[16,17], Fred H. Gage[7,18], Ursula Bellugi[12], Eric Halgren[8,9,18], Katerina Semendeferi[5,18,19] & Alysson R. Muotri[1,2,3,18,19]

Williams syndrome is a genetic neurodevelopmental disorder characterized by an uncommon hypersociability and a mosaic of retained and compromised linguistic and cognitive abilities. Nearly all clinically diagnosed individuals with Williams syndrome lack precisely the same set of genes, with breakpoints in chromosome band 7q11.23 (refs 1–5). The contribution of specific genes to the neuroanatomical and functional alterations, leading to behavioural pathologies in humans, remains largely unexplored. Here we investigate neural progenitor cells and cortical neurons derived from Williams syndrome and typically developing induced pluripotent stem cells. Neural progenitor cells in Williams syndrome have an increased doubling time and apoptosis compared with typically developing neural progenitor cells. Using an individual with atypical Williams syndrome[6,7], we narrowed this cellular phenotype to a single gene candidate, frizzled 9 (*FZD9*). At the neuronal stage, layer V/VI cortical neurons derived from Williams syndrome were characterized by longer total dendrites, increased numbers of spines and synapses, aberrant calcium oscillation and altered network connectivity. Morphometric alterations observed in neurons from Williams syndrome were validated after Golgi staining of post-mortem layer V/VI cortical neurons. This model of human induced pluripotent stem cells[8] fills the current knowledge gap in the cellular biology of Williams syndrome and could lead to further insights into the molecular mechanism underlying the disorder and the human social brain.

This study included participants with a clinical diagnosis of Williams syndrome (WS): individuals harbouring typical gene deletions in the Williams–Beuren syndrome critical region[1,9] (WS17, 25, 77 and 79) and an individual with atypical WS with a partial deletion (pWS88) as well as typically developing (TD) participants (TD55, 59, 63 and 70) (Fig. 1a and Extended Data Fig. 1a). After a series of cognitive and social profiles[2,5,10], we confirmed that the individuals with typically deleted WS were a representative cohort of the disorder (Fig. 1b–c, Extended Data Fig. 1b–g and Supplementary Note 1). To generate a human cellular model of WS[11,12], dental pulp cells (DPCs) obtained from participants' deciduous teeth were reprogrammed into induced pluripotent stem cells (iPSCs) (Extended Data Fig. 2a). We selected two to three clones from each individual for further investigation (Extended Data Fig. 2b–g and Supplementary Tables 1 and 2). To obtain the relevant cells, iPSC clones underwent neural induction (Fig. 1d) and neural progenitor cells (NPCs) were further characterized (Fig. 1e–g). Cortical neurons were obtained using a modified protocol from our previous publication[11] (Fig. 1h–j). Finally, iPSC-derived neurons exhibited a complete set of electrophysiological properties (Fig. 1k–m and Extended Data Fig. 2h, i).

The impact of the genome-wide Williams–Beuren syndrome chromosome region deletion was determined by unbiased RNA sequencing (RNA-seq) (Extended Data Fig. 3a–d). Differential expression analyses revealed misregulated genes among the three genotypes (Extended Data Fig. 3e–h, Extended Data Table 1 and Supplementary Tables 3–9). Gene ontology (GO) analyses of NPCs and neurons revealed biological processes relevant to the condition (Extended Data Fig. 3h, i, Extended Data Table 2 and 3 and Supplementary Table 10). Remarkably, 'cell adhesion', 'axon guidance' and 'cell maturation' were also among the top-ranking categories detected in an independent publication using WS NPC gene expression analysis[13].

As suggested by the NPC global gene expression analyses, during the culture maintenance, typical WS NPCs became confluent more slowly than TD NPCs (Fig. 2a). After plating the same number of NPCs, we verified that the number of typical WS NPCs on day 4 was less than the TD NPCs (Fig. 2b and Extended Data Fig. 4a). To rule out the possibility that the difference in heterogeneity of iPSC-derived NPCs could result in this observation, the NPC population was fully characterized and no difference between WS and TDs were found (Fig. 1e–g and Extended Data Fig. 4b). We also used single-cell gene expression profiling to access the homogeneity of the NPCs (Fig. 2c–e and Extended Data Fig. 4c–g). We further investigated the proliferation of WS NPCs by performing BrdU labelling, immunostaining and fluorescence-activated cell sorting (FACS) (Extended Data Fig. 4h, i). Since no difference was found, we assessed apoptosis in WS NPCs using DNA fragmentation (propidium iodide) and caspase assay (Extended Data Fig. 4j). We found a significant increase in subG1

[1]University of California San Diego, School of Medicine, UCSD Stem Cell Program, Department of Pediatrics/Rady Children's Hospital San Diego, La Jolla, California 92037, USA. [2]University of California San Diego, School of Medicine, Department of Cellular & Molecular Medicine, La Jolla, California 92037, USA. [3]Center for Academic Research and Training in Anthropogeny (CARTA), La Jolla, California 92093, USA. [4]National Center for Genetic Engineering and Biotechnology (BIOTEC), Virology and Cell Technology Laboratory, Pathum Thani 12120, Thailand. [5]University of California San Diego, Department of Anthropology, La Jolla, California 92093, USA. [6]Graduate Program in Health Sciences, School of Medicine, Pontifícia Universidade Católica do Paraná (PUCPR), Curitiba, Paraná, Brazil. [7]The Salk Institute for Biological Studies, Laboratory of Genetics, La Jolla, California 92037, USA. [8]University of California San Diego, Multimodal Imaging Laboratory, La Jolla, California 92093, USA. [9]University of California San Diego, School of Medicine, Department of Neurosciences, La Jolla, California 92093, USA. [10]University of California San Diego, Center for Human Development, La Jolla, California 92093, USA. [11]SAHMRI Mind & Brain Theme, Laboratory for Human Neurophysiology and Genetics, Flinders University School of Medicine, Adelaide, South Australia 5000, Australia. [12]The Salk Institute for Biological Studies, Laboratory for Cognitive Neuroscience, La Jolla, California 92037, USA. [13]Colorado College, Department of Psychology, Colorado Springs, Colorado 80903, USA. [14]University of California San Diego, School of Medicine, Department of Radiology, La Jolla, California 92093, USA. [15]University of California San Diego, Department of Cognitive Science, La Jolla, California 92093, USA. [16]University of Utah, Department of Pediatrics, Salt Lake City, Utah 84108, USA. [17]University of Utah, The Brain Institute, Salt Lake City, Utah 84108, USA. [18]University of California San Diego, Kavli Institute for Brain and Mind, La Jolla, California 92093, USA. [19]University of California San Diego, School of Medicine, Neuroscience Graduate Program, La Jolla, California 92093, USA.

*These authors contributed equally to this work.

‡Deceased.

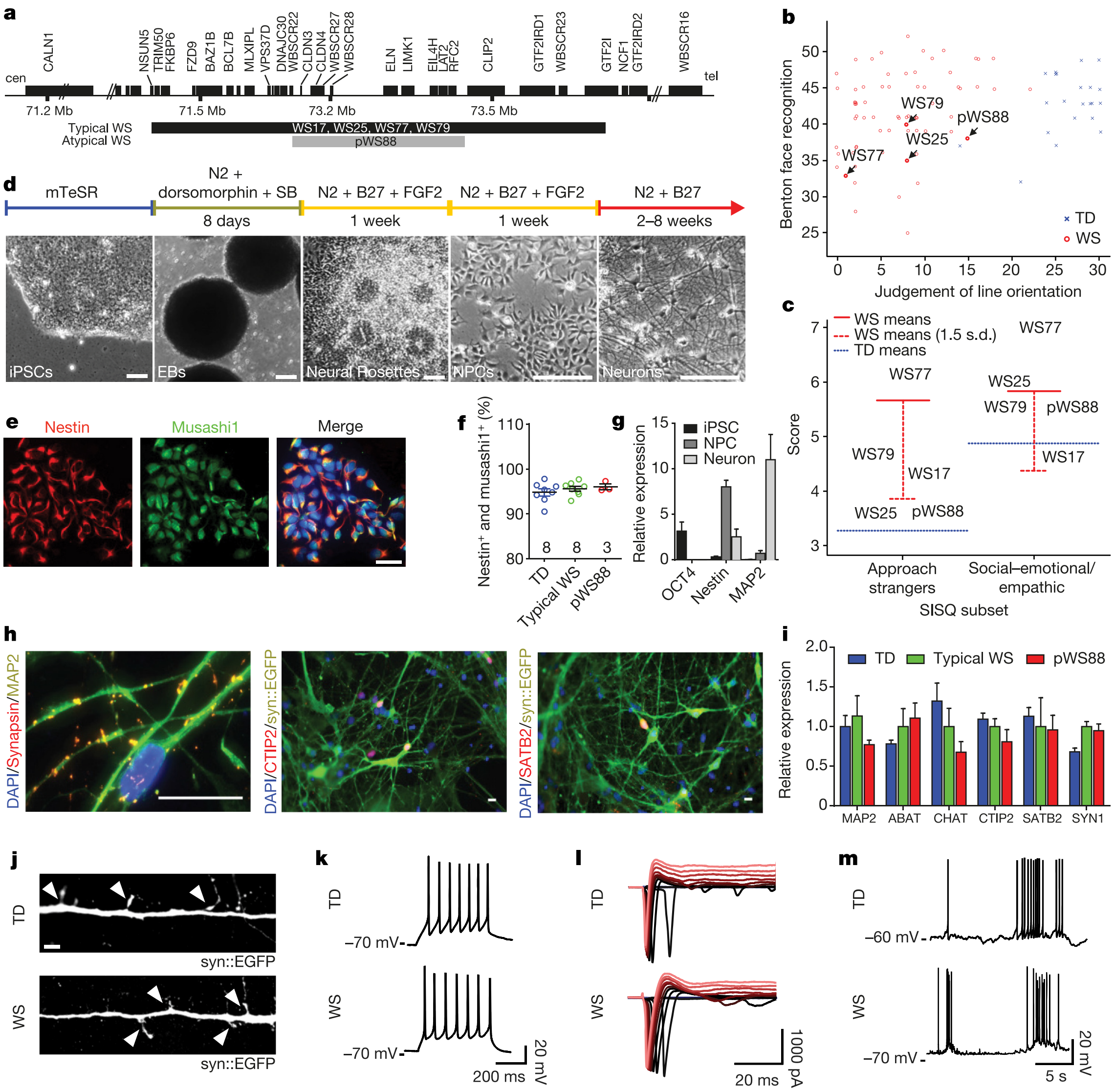


Figure 1 | Characterization of participating individuals and iPSC differentiation. a, Diagram showing genes and deletion region of individuals with WS. **b**, Scatter plot of Benton Face Recognition and Judgment of Line Orientation scores (jitter added) for $n = 69$ individuals with WS and $n = 22$ TD participants. **c**, Mean test scores for WS (solid red lines; $n = 101$ for approach strangers; $n = 100$ for social–emotional/empathic) and for TD individuals (dotted blue lines; $n = 80$ for approach strangers; $n = 79$ for social–emotional/empathic). **d**, Neural induction and neuronal differentiation protocol. Scale bar, 50 μm. **e**, Stage-specific protein expression in iPSC-derived NPCs. Scale bar, 50 μm. **f**, High percentage of Nestin and Musashi1-positive population was comparably observed in TD, typical WS and pWS88 NPCs by FACS. Data are shown as mean ± s.e.m.; *n*, number of clones. **g**, Stage-specific markers for iPSC (*OCT4*), NPC (Nestin) and neuron (*MAP2*) by quantitative PCR (qPCR). **h**, Stage-specific protein expression in 6-week-old neurons. Scale bar, 25 μm. **i**, Expression of different neuronal markers in neurons indicating multiple neuronal subtypes in 6-week-old culture by qPCR. Data are shown as mean ± s.e.m. **j**, A representative image of neuronal protrusions (spine-like, arrowheads) from iPSC-derived neurons. Scale bar, 2 μm. **k**–**m**, Four-week-old TD and WS iPSC-derived neurons show evoked action potentials (**k**), evoked voltage-dependent sodium and potassium currents (**l**) and spontaneous bursts of action potentials (**m**).

(Fig. 2f, g) and caspase-positive populations (Fig. 2h, i) in WS NPCs, indicating increased apoptosis.

FZD9 is expressed in NPCs[14] (Fig. 2j) and has been shown to regulate cell division and programmed cell death in different cell types[15,16]. In our study, *FZD9* was hemizygously deleted in the participants with typical WS, but retained in atypical pWS88 (Fig. 1a and Extended Data Fig. 1b). Thus, we hypothesized that *FZD9* regulates human NPC apoptosis. We transduced TD NPCs with a lentivirus carrying either short hairpin RNA (shRNA) against *FZD9* (shFZD9) or non-specific shRNA (shControl) and WS NPCs with lentiviruses carrying a *FZD9* cDNA construct (Fig. 2k, l). TD NPCs transduced with shFZD9 showed a reduction in the number of cells on day 4 (Fig. 2m) and an increase in the subG1 population (Fig. 2n) and caspase activity (Fig. 2o) compared with TD NPCs expressing the shControl. Similar results were observed in atypical pWS88 (Extended Data Fig. 4k–n). Restoring *FZD9* expression in typical WS NPCs brought the number of NPCs on day 4/day 0 to a similar level to TD NPCs. It also significantly reduced the apoptotic population to the TD level.

Since several Wnt genes were downregulated in WS NPCs (Extended Data Tables 2 and 3) and FZD9 can be activated by Wnt ligands, we tested if we could rescue the NPC viability by treating cells with the GSK3 inhibitor CHIR98014 (ref. 17). First, we confirmed that the canonical Wnt pathway was affected in WS by measuring the Axin2 and SP5 expression levels, two universal Wnt target genes[18,19]. Both genes were significantly downregulated in WS cells compared with TDs (Fig. 2p, q). By treating WS NPCs with CHIR98014, we were able to rescue cell viability (Fig. 2r). Together, our results indicate a role for FZD9 in NPC viability.

Our protocol generated a consistent population of forebrain neurons, confirmed by the pan-neuronal and subtype-specific cortical markers

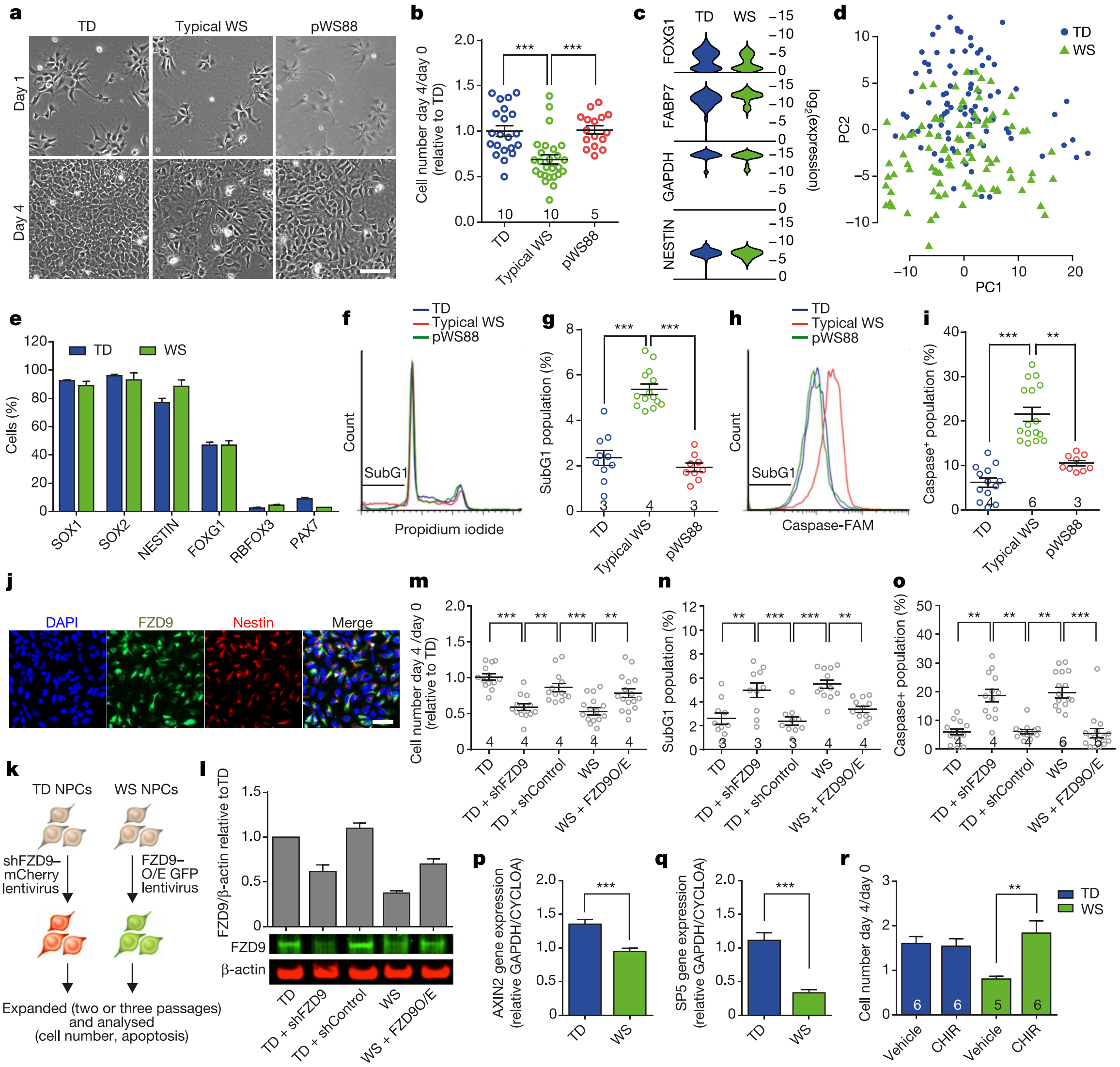


Figure 2 | Defect in apoptosis of WS-derived NPCs owing to haploinsufficiency of *FZD9*. **a**, Representative images showing the difference in confluency between TD, typical WS and pWS88 iPSC-derived NPCs on day 4. Scale bar, 100 μm. **b**, Ratio of NPC number on day 4 over day 0 relative to TD. **c**, Violin plots of representative genes expressed in NPCs from single-cell analyses. **d**, Principal component analysis (PCA) was used to compare the expression levels in individual cells on the basis of the first two principal components. **e**, Percentage of cells expressing NPC, neuronal (*RBFOX3*) and neural crest (*PAX7*, contaminant population) related genes. WS and TD iPSC-derived NPCs show similar percentages of cells expressing target genes over defined cycle threshold (C_t) control values. **f**, Representative propidium iodide histogram showing an increase in subG1 population in typical WS NPCs. **g**, Percentage of subG1 population. **h**, Representative histogram showing an increase in caspase activity (caspase-FAM intensity) in typical WS NPCs. **i**, Percentage of population with high caspase activity. **j**, FZD9 protein expression in TD iPSC-derived NPCs. **k**, Schematic of *FZD9* gain/loss of function experiments in NPCs. **l**, Expression level of FZD9 protein after treatment with shFZD9, shControl and *FZD9* overexpression vectors, assessed by western blot analysis. **m**–**o**, Ratio of NPC number on day 4 over day 0 relative to TD (**m**), percentage of subG1 population (**n**) and percentage of population with high caspase activity (**o**) when TD NPCs were treated with shFZD9 and shControl, and WS NPCs were overexpressed with *FZD9*. **p**–**q**, Significant decrease in expression of Axin2 (**p**) and SP5 (**q**) of WS NPCs compared with TDs. **r**, Rescue of WS NPC viability after CHIR98014 treatment. All data are shown as mean ± s.e.m.; *n*, number of clones. $^{**}P < 0.01$, $^{***}P < 0.001$, Kruskal–Wallis test and Dunn's multiple comparison test (**b**, **g**, **i**), one-way ANOVA and Tukey's post hoc test (**g**, **m**–**o**), two-sided unpaired Student's *t* test (**p**, **q**), two-sided unpaired Mann–Whitney test (**r**).

such as *CTIP2* (layers V/VI)[20–22] and *SATB2* (layer III) (Fig. 1h, i). The neuronal population was also characterized by single-cell gene expression profiling (Fig. 3a–c and Extended Data Fig. 5a–f), revealing mostly glutamatergic neurons, with a small population of GABAergic (γ-aminobutyric-acid-releasing) neurons and glia (Fig. 3a). We did not detect significant variability in these subtypes of neurons expressing target genes or in the expression levels of several markers for cortical layers and neurotransmitters among the genotypes. However, we did detect differences in the expression of specific genes in these populations (Fig. 3b, c) that could lead to specific alterations in mature neurons. We focused specifically on markers for cortical layers V/VI, since pathologies affecting these layers have been reported in disorders with compromised social functioning, such as autism[23]. We found that typical WS iPSC-derived CTIP2-positive neurons had significantly higher total dendritic length, dendrite number and number of dendritic spines than TDs (Fig. 3d–g and Extended Data Fig. 6a–m). Interestingly, atypical pWS88 neurons were morphologically similar to TD neurons except for dendrite number (Fig. 3f). To determine whether the WS neuronal phenotype was cell autonomous or dependent on other cells or culture conditions, we recorded the dendritic growth over time. The result showed a faster dendritic growth rate in WS neurons compared with TD or pWS88 (Extended Data Fig. 6n–r).

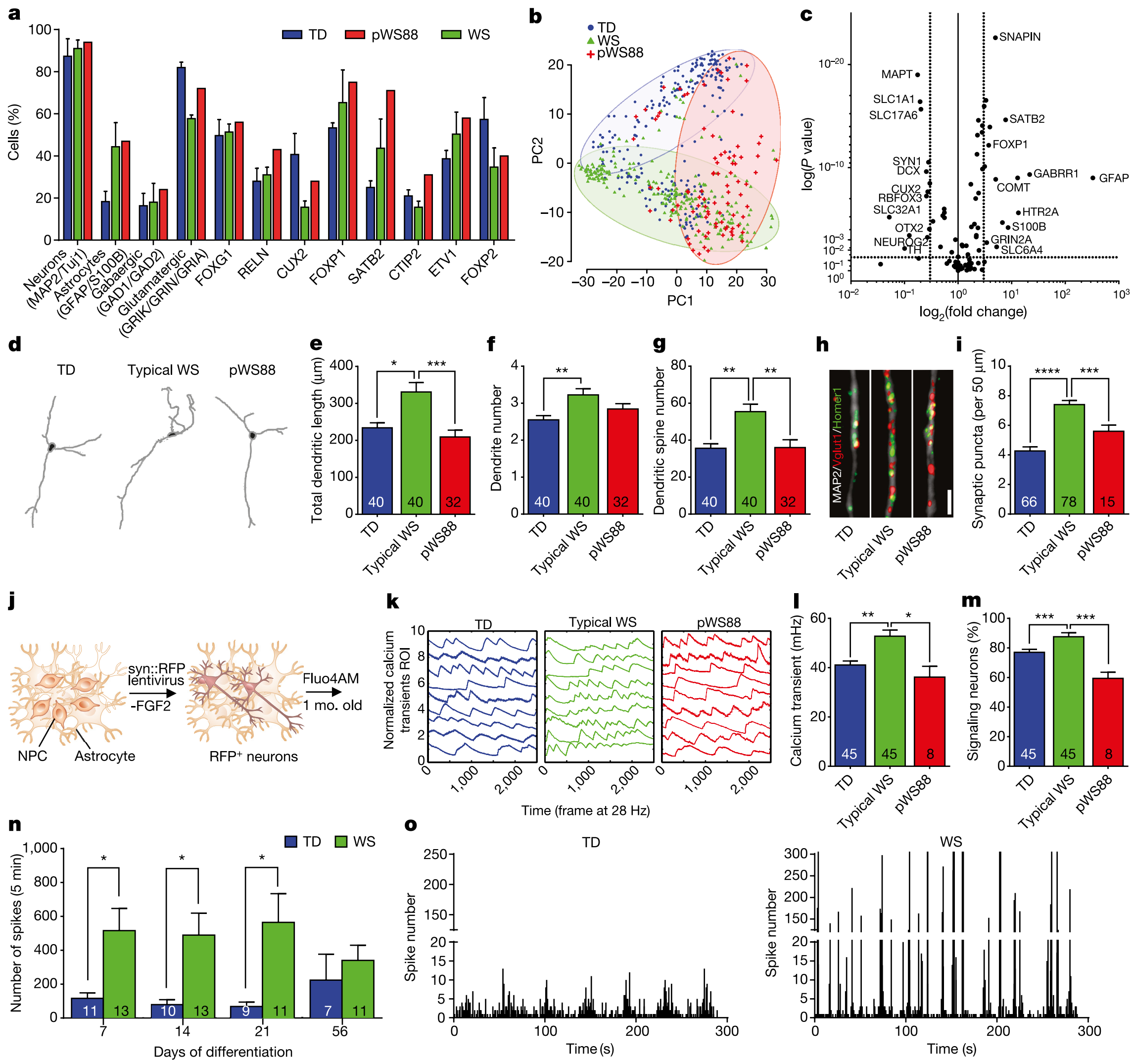


Figure 3 | Altered morphology of WS-derived cortical neurons and network activity. a, Percentage of cells expressing neural markers, neurotransmitter and cortical layer-related genes. WS, pWS88 and TD iPSC-derived neurons show non-significant percentage of cells expressing target genes over defined control C_t value. **b**, PCA of 672 cells projected onto the first two components. Overlaid populations of TD, pWS88 and WS neurons are shown. **c**, Volcano plot illustrates differences in expression patterns of target genes of iPSC-derived neurons from the single-cell analyses. The dotted lines represent more than or equal to 3.0-fold differentially expressed genes between the groups at $P < 0.05$ (unpaired Student's *t* test). **d**, Representative images of tracings from TD, typical WS and atypical pWS88 iPSC-derived neurons (Syn::EGFP- and CTIP2-positive neurons). **e**–**g**, Morphometric analyses showing significant differences between TD, typical WS and pWS88 in total dendritic length (**e**), between TD and typical WS in dendrite number (**f**) and between TD, typical WS and pWS88 in dendritic spine number (**g**). **h**, **i**, Puncta quantification of post- and presynaptic markers. Scale bar, 2 μm. For **e**–**g** and **i**, data are shown as mean ± s.e.m.; *n*, number of traced neurons. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, $^{****}P < 0.0001$, Kruskal–Wallis test and Dunn's multiple comparison test (**e**–**g**), one-way ANOVA and Tukey's post hoc test (**i**). **j**, Schematic diagram summarizing preparation of neurons for calcium transient analysis. **k**, Representative images of the calcium tracing from iPSC-derived neurons. Fluorescence intensity changes reflecting intracellular calcium fluctuations in neurons in different regions of interest (ROI). **l**, **m**, Typical WS-derived neurons exhibited significant increase in calcium transient frequency (**l**) and percentage of signalling neuron in the culture (**m**) when compared with TD or pWS88 neurons. Data are shown as mean ± s.e.m.; *n*, number of fields analysed; 3,198 neurons for TD, 4,446 neurons for WS and 48 neurons for pWS88. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, Kruskal–Wallis test and Dunn's multiple comparison test. **n**, MEA analyses revealed an increase in spontaneous neuronal spikes in WS during differentiation compared with TD. **o**, Although the number of total network bursts do not differ, WS shows a higher number of spikes in each burst compared with TD. Data are shown as mean ± s.e.m.; *n*, number of MEA wells analysed. $^{*}P < 0.05$, $^{**}P < 0.01$, two-sided unpaired Student's *t* test.

Also, no differences were observed in the total dendritic length, segment number or spine density using NPCs plated at different cellular densities (Extended Data Fig. 6 s–u).

An increase in the number of dendritic spines per neuron could lead to an increase in synaptic contacts and, therefore, synaptic activity[24], which could result in functional alterations. WS neurons had significantly more glutamatergic excitatory synapses compared with TD and pWS88 (Fig. 3h, i and Extended Data Figs 6v and 7a) and an increased frequency of calcium transients with a higher percentage of signalling neurons in WS cultures (Fig. 3j–m and Extended Data Fig. 7b–f). Using multi-electrode array (MEA) electrophysiology, our data showed that WS neuronal cultures had a significant increase in spike frequency compared with TD-derived neurons (Fig. 3n, o and Extended Data Fig. 7g, h).

In an attempt to place our iPSC findings in the larger context of the cortical morphology of human participants at the gross anatomical and cellular levels, we conducted two sets of additional experiments to test predictions based on NPC and neuronal differences found *in vitro*. In addition to the total volume reduction in WS brains previously reported[25], multivariate analyses of variance (MANOVA) from

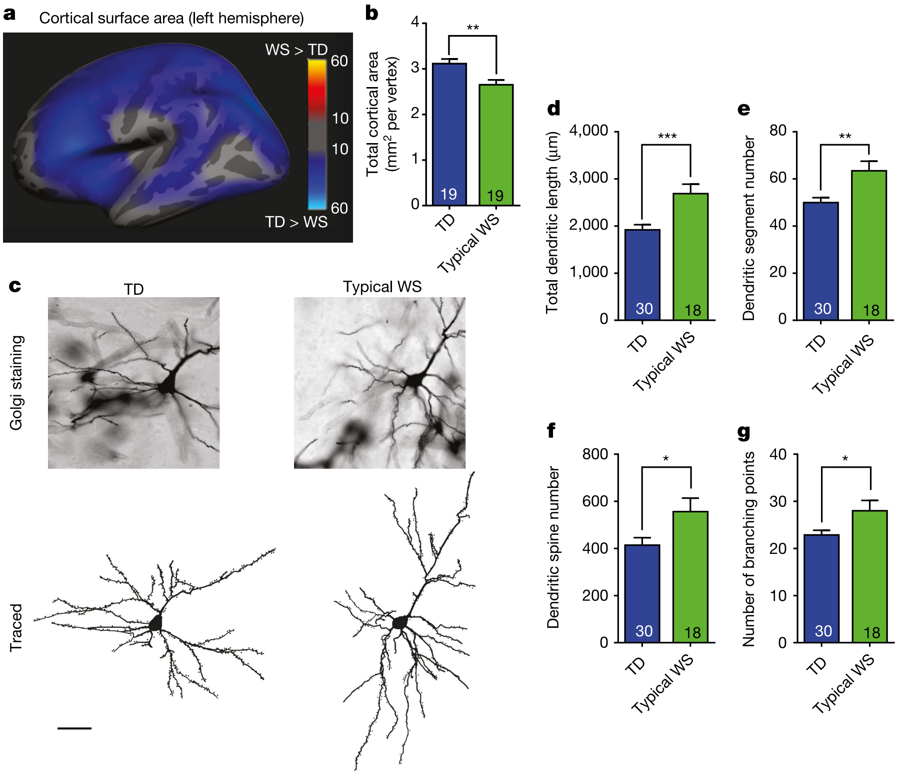


Figure 4 | Neuroanatomical and morphological alterations in WS human brains. a, Statistical parametric map of the vertex-wise group differences between TD and WS in cortical surface area (left hemisphere shown) assessed by structural MRI scans. Colour scales indicate the *P* value for statistical test: blue, decrease; grey, no difference. Statistics are displayed on a template group-averaged cortical surface rendering of TD adult participants. **b**, Reduction in overall cerebral cortical surface area in WS. Data are shown as mean ± s.e.m.; *n*, number of brains analysed. $^{**}P < 0.01$, one-sided unpaired Student's *t* test. **c**, Representative images of post-mortem cortical layer V/VI pyramidal neurons using Golgi staining (top) and their corresponding tracing (bottom) from TD and WS. **d–g**, Morphometric analysis showing significant increases in total dendritic length (**d**), dendritic spine numbers (**e**), dendritic segment number (**f**) and number of branching points (**g**) in WS compared with TD post-mortem cortical layer V/VI pyramidal neurons. Data are shown as mean ± s.e.m.; *n*, number of traced neurons. $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$, two-sided unpaired Student's *t* test (**d**), two-sided unpaired Mann–Whitney test (**e**–**g**).

our structural brain imaging of living participants revealed a significant decrease in overall cortical surface area in WS compared with TD individuals (Fig. 4a, b), but not in cortical thickness. Additionally, we conducted a separate set of experiments on post-mortem brains from WS and TD tissue donors to investigate possible alterations in the morphology of cortical neurons, as predicted by our iPSC findings. Similar to WS iPSC-derived CTIP2-positive neurons, post-mortem layer V/VI pyramidal neurons displayed larger total dendritic length and higher number of dendritic spines (Fig. 4c–e), and similar spine density and soma area, than TD neurons (Extended Data Fig. 7i–s). Post-mortem layer V/VI neurons also showed significantly increased numbers of dendritic segments and branching points (Fig. 4f, g), and similar numbers of dendritic trees.

The morphometric data in combination with the increased glutamatergic gene expression and number of co-localized synaptic puncta observed in neurons from WS suggest that an increased number of synapses may result in the altered network activity, which could contribute to the characteristic behaviour of individuals with WS. Our study reveals that the WS phenotypes described here are the foundation for the understanding of the complex human social behaviour. This approach provides an additional strategy to study the cellular and molecular underpinnings of complex human attributes, such as language in a social environment.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 30 December 2014; accepted 29 June 2016.
Published online 10 August 2016.

1. Korenberg, J. R. *et al.* VI. Genome structure and cognitive map of Williams syndrome. *J. Cogn. Neurosci.* **12** (Suppl. 1), 89–107 (2000).
2. Meyer-Lindenberg, A. *et al.* Neural basis of genetically determined visuospatial construction deficit in Williams syndrome. *Neuron* **43,** 623–631 (2004).
3. Bellugi, U., Lichtenberger, L., Mills, D., Galaburda, A. & Korenberg, J. R. Bridging cognition, the brain and molecular genetics: evidence from Williams syndrome. *Trends Neurosci*. **22,** 197–207 (1999).
4. Bellugi, U., Lichtenberger, L., Jones, W., Lai, Z. & St George, M. I. The neurocognitive profile of Williams syndrome: a complex pattern of strengths and weaknesses. *J. Cogn. Neurosci.* **12** (Suppl. 1), 7–29 (2000).
5. Doyle, T. F., Bellugi, U., Korenberg, J. R. & Graham, J. "Everybody in the world is my friend" hypersociability in young children with Williams syndrome. *Am. J. Med. Genet. A* **124,** 263–273 (2004).
6. Dai, L. *et al.* Is it Williams syndrome? GTF2IRD1 implicated in visual-spatial construction and GTF2I in sociability revealed by high resolution arrays. *Am. J. Med. Genet. A* **149A,** 302–314 (2009).
7. Edelmann, L. *et al.* An atypical deletion of the Williams–Beuren syndrome interval implicates genes associated with defective visuospatial processing and autism. *J. Med. Genet.* **44,** 136–143 (2007).
8. Chailangkarn, T., Acab, A. & Muotri, A. R. Modeling neurodevelopmental disorders using human neurons. *Curr. Opin. Neurobiol.* **22,** 785–790 (2012).
9. Ewart, A. K. *et al.* Hemizygosity at the elastin locus in a developmental disorder, Williams syndrome. *Nature Genet.* **5,** 11–16 (1993).
10. Järvinen-Pasley, A. *et al.* Defining the social phenotype in Williams syndrome: a model for linking gene, the brain, and behavior. *Dev. Psychopathol.* **20,** 1–35 (2008).
11. Marchetto, M. C. *et al.* A model for neural development and treatment of Rett syndrome using human induced pluripotent stem cells. *Cell* **143,** 527–539 (2010).
12. Beltrão-Braga, P. C. B. *et al.* Feeder-free derivation of induced pluripotent stem cells from human immature dental pulp stem cells. *Cell Transplant.* **20,** 1707–1719 (2011).
13. Adamo, A. *et al.* 7q11.23 dosage-dependent dysregulation in human pluripotent stem cells affects transcriptional programs in disease-relevant lineages. *Nature Genet.* **47,** 132–141 (2015).
14. Van Raay, T. J. *et al.* frizzled 9 is expressed in neural precursor cells in the developing neural tube. *Dev. Genes Evol.* **211,** 453–457 (2001).

15. Zhao, C. *et al.* Hippocampal and visuospatial learning defects in mice with a deletion of frizzled 9, a gene in the Williams syndrome deletion interval. *Development* **132,** 2917–2927 (2005).
16. Fujimoto, T., Tomizawa, M. & Yokosuka, O. SiRNA of frizzled-9 suppresses proliferation and motility of hepatoma cells. *Int. J. Oncol.* **35,** 861–866 (2009).
17. Lian, X. *et al.* Efficient differentiation of human pluripotent stem cells to endothelial progenitors via small-molecule activation of WNT signaling. *Stem Cell Rep*. **3,** 804–816 (2014).
18. Jho, E. H. *et al.* Wnt/beta-catenin/Tcf signaling induces the transcription of Axin2, a negative regulator of the signaling pathway. *Mol. Cell. Biol.* **22,** 1172–1183 (2002).
19. Fujimura, N. *et al.* Wnt-mediated down-regulation of Sp1 target genes by a transcriptional repressor Sp5. *J. Biol. Chem.* **282,** 1225–1237 (2007).
20. Srinivasan, K. *et al.* A network of genetic repression and derepression specifies projection fates in the developing neocortex. *Proc. Natl Acad. Sci. USA* **109,** 19071–19078 (2012).
21. Chen, B. *et al.* The Fezf2–Ctip2 genetic pathway regulates the fate choice of subcortical projection neurons in the developing cerebral cortex. *Proc. Natl Acad. Sci. USA* **105,** 11382–11387 (2008).
22. Leone, D. P., Srinivasan, K., Chen, B., Alcamo, E. & McConnell, S. K. The determination of projection neuron identity in the developing cerebral cortex. *Curr. Opin. Neurobiol.* **18,** 28–35 (2008).
23. Hutsler, J. J. & Zhang, H. Increased dendritic spine densities on cortical projection neurons in autism spectrum disorders. *Brain Res.* **1309,** 83–94 (2010).
24. Spitzer, N. C., Root, C. M. & Borodinsky, L. N. Orchestrating neuronal differentiation: patterns of Ca^{2+} spikes specify transmitter choice. *Trends Neurosci*. **27,** 415–421 (2004).
25. Chiang, M. C. *et al.* 3D pattern of brain abnormalities in Williams syndrome visualized using tensor-based morphometry. *Neuroimage* **36,** 1096–1109 (2007).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was supported by grants from the California Institute for Regenerative Medicine (CIRM) TR2-01814 and TR4-06747, the National Institutes of Health (NIH) through P01 NICHD033113, NIH Director’s New Innovator Award Program 1-DP2-OD006495-01, R01MH094753, R01MH103134, U19MH107367, U19MH106434, R01MH095741, a National Alliance for Research on Schizophrenia and Depression (NARSAD) Independent Investigator Grant to A.R.M., grants from the Bob and Mary Jane Engman, the JPB Foundation, Paul G. Allen Family Foundation, the Leona M. and Harry B. Helmsley Charitable Trust grant 2012-PG-MED002, Annette C. Merle-Smith, the G. Harold & Leila Y. Mathers Foundation, the Royal Thai Government Scholarship to T.C., a CIRM postdoctoral fellowship to C.A.T., the Rita L. Atkinson Graduate fellowship to B.H.-M and the University of California San Diego Kavli Institute for Brain and Mind. Human tissue was obtained from the University of Maryland Brain and Tissue Bank, which is a brain and tissue repository of the NIH NeuroBioBank. We acknowledge K. Jepsen for the DNA bead arrays and members of the Willert laboratory for assistance with the Wnt pathway experiments. We thank all the participants and their families.

Author Contributions A.R.M., T.C. and C.A.T. designed the experiments and wrote the manuscript with input from K.S. and all authors. T.C. processed DPCs, generated and characterized iPSCs, NPCs and neurons, and performed cell number, proliferation, and apoptosis experiments as well as *FZD9* knockdown and overexpression and statistical analysis. C.A.T. performed C1 single-cell analyses, synaptic quantification, calcium imaging, cell density experiments, live neuronal morphology analysis and statistical analysis. B.C.F. performed MEA recording, PCR for retrovirus silencing and Wnt pathway gene-expression analysis. B.C.F. and S.E.R. prepared astrocytes for co-culture experiments, NPC characterization by flow cytometry and CHIR 98014 experiments. K.S. designed all morphometry experiments with B.H.-M. and B.J., and co-wrote the manuscript to link the various levels of investigation from the whole-brain imaging findings to the cellular level. L.S. prepared Golgi staining for post-mortem neurons with help from K.L.H. and B.J. B.H.-M. obtained morphometric data on iPSC-derived neurons and post-mortem neurons. D.X.Y., M.C.M., C.A.T. and L.M. performed calcium transient experiments and statistical analysis. T.T.B. performed brain scan and statistical analysis with help from A.M.D. C.B. performed electrophysiological tests. M.D., W.W., P.L. and Y.M.S. performed neurocognitive and social tests. A.J., Y.M.S., and M.C.A. performed analyses and interpretation of social/neurocognitive tests. R.H.H. performed bioinformatics analysis. L.D. and J.R.K. confirmed deletion of all cells from participants with WS who donated them for reprogramming. E.H., U.B., F.H.G., K.S. and A.R.M. edited the manuscript for publication.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to A.R.M. (muotri@ucsd.edu) or K.S. (ksemende@ucsd.edu).

METHODS

Participants for behavioural study and source of cells for reprogramming. The study protocols were approved by University of California San Diego and Salk Institute IRB/ESCRO committees (protocols 141223ZF and 95-0001, respectively). Four TD individuals (ages 8–19 years) and five individuals with WS (ages 8–14 years; Extended Data Fig. 1a) were included in the analysis: four of the latter group had typical WS gene deletions and one (pWS88) had a partial deletion in the WS region. Informed consents were obtained from all participants or their parents as appropriate. Genetic diagnosis of WS was established using fluorescent *in situ* hybridization probes for elastin (*ELN*), a gene consistently associated with the deletion in the typical WS region[1,9]. All of the participants with WS having confirmed genetic deletion exhibited the medical and clinical characteristics of the WS phenotype, including previously established cognitive, behavioural and physical features associated with the syndrome[4]. A diagnosis of WS was confirmed on the basis of the Diagnostic Score Sheet (DSS) for WS (American Academy of Paediatrics Committee on Genetics, 2001), with a particular focus on the cardiovascular abnormalities and the characteristic facial features associated with the *ELN* deletion. The scores for the participants were at the mean for WS (9) or higher, with the individual with partial deletion in the WS chromosomal region (pWS88) scoring lower than the individuals with typical WS deletion. Similarly, pWS88 reported fewer symptoms with connective tissue and growth, his cognitive scores were slightly higher than the typical individuals with WS, and he did not demonstrate the disparity between verbal and visual–spatial abilities typical of WS. However, pWS88 did display behavioural and developmental features consistent with WS, including developmental delay, over-friendliness and anxiousness.

Behavioural and neurocognitive tests. The participants were administered standard tests to quantify their non-verbal and verbal abilities, as well as versions of the WS cognitive and social profiles to capture the distinct pattern of strengths and weaknesses both within and across domains associated with the WS cognitive and social phenotype. Details of the tests and the measures tapping into the two profiles are presented in Extended Data Fig. 1. The WS cognitive profiles for the five participants with WS were constructed by calculating the log of predictive likelihood ratios under assumed normality for age-appropriate TD versus WS classifications on the basis of verbal and performance IQ (VIQ and PIQ), Beery Developmental Test of Visual-Motor Integration (VMI) and Peabody Picture Vocabulary Test (PPVT) standard scores, subject to availability. Predictive distributions were based on the published normative mean and s.d. for each of the tests employed, whereas for the WS classification the predictive distributions[26] were determined using data from $n = 81$ (VIQ and PIQ), $n = 56$ (VMI) and $n = 97$ (PPVT) participants in a broader WS sample (described in Extended Data Fig. 1d). A tobit model was used to estimate parameters for individuals with WS on the VMI owing to the presence of floor effects. The WS social profiles for the five participants with WS were constructed using measures of social approach behaviour, emotionality/empathy and language use.

Confirmation of WS deletion. Quantitative PCR was used to define the breakpoints of deleted regions in DNA isolated from iPSCs, or lymphoblast cell lines for participants with WS, with probes spanning from CALN1 to WBSCR16 and template DNA. Taqman expression assay probes detecting the WS region genes were designed and synthesized with sequences shown in Supplementary Table 11. RNase P (VIC) was used as control. Quantitative PCR was performed on the ABI PRISM 7900HT system and the results were analysed using SDS 3.2.

Cell collection, reprogramming and characterization. We avoided invasive sample collection methods such as skin biopsy or blood withdrawal by taking advantage of the natural loss of deciduous teeth as a source of somatic cells. We chose to reprogram dental pulp cells (DPCs) because these cells develop from the same set of early progenitors that generate neurons. Furthermore, the neurons derived from iPSCs generated from DPCs express higher levels of forebrain genes compared with those generated from skin fibroblast-derived iPSCs[27], serving the purpose of this study. Deciduous teeth were collected when they fell out and were shipped to our laboratory in DMEM 1× (Mediatech) with 4% Pen/Strep (Mediatech). Dental pulp was pulled out, washed in PBS with 4% Pen/Strep and incubated in 5% TrypLE (Gibco) for 15 min. Pulp was partly dissociated using needles and plated in culture medium (DMEM/F12 50:50, 15% FBS, 1%NEAA, 1% fungizone and 2% Pen/Strep). In 1–4 weeks, DPCs migrated out of the pulp and could be passaged and frozen as stock. DPCs in early passage (two to three) were reprogrammed using pMXs retroviruses expressing Yamanaka transcription factors (obtained from Addgene, Cambridge, Massachusetts)[12]. After 4 days, transduced DPCs were trypsinized, plated on mouse embryonic fibroblasts and cultured using human embryonic stem cell (hESC) medium. After manually picked and clonally expanded, feeder-free iPSCs were grown on matrigel-coated dishes (BD Bioscience, San Jose, California) with mTeSR1 (StemCell Technologies) or iDEAL[28].

Karyotyping. All G-banding karyotyping analyses were performed by Molecular Diagnostics Service (San Diego, California) and Children's Hospital Los Angeles (Los Angeles, California).

Genotyping. Two hundred nanograms of DNA were processed and hybridized to the Illumina Infinium Human Core Exome BeadChip following manufacturer's instructions. Illumina GenomeStudio V2011.1 with the Genotyping Module version 1.9.4 was used to normalize data and call genotypes using reference data provided by Illumina. Illumina's cnv Partition and gada R packages were used to automatically detect aberrant copy number region. In addition, the B Allele Frequency (BAF) and Log R Ratio (LRR) distributions were manually checked to determine additional CNVs not detected by the software. Sample identification/relatedness was assessed by comparing called genotypes for each sample. The absolute number of different genotypes was counted and the Euclidean distances were calculated to identify relatedness of the samples.

Teratoma assay. Dissociated iPSC colonies were centrifuged and resuspended in 1:1 matrigel and phosphate buffer saline solution. The cells were injected subcutaneously in nude mice. After 1–2 months, teratomas were dissected, fixed and sliced. Sections were stained with haematoxylin and eosin for further analysis. Protocols were previously approved by the University of California San Diego Institutional Animal Care and Use Committee.

Neural induction and neuronal differentiation. iPSCs were cultured on matrigel-coated dishes and fed daily with mTeSR for 7 days. On the next day, mTeSR was substituted by N2 medium (DMEM/F12 supplemented with 0.5× N2-Supplement (Life Technologies), 1 μM dorsomorphin (Tocris) and 1 μM SB431542 (Stemgent)) for 1–2 days. iPSC colonies were lifted off, cultured in suspension on the shaker (95 r.p.m. at 37 °C) for 8 days to form embryoid bodies and fed with N2 media. Embryoid bodies then were mechanically dissociated, plated on a matrigel-coated dish and fed with N2B27 medium (DMEM/F12 supplemented with 0.5× N2-Supplement, 0.5× B27-Supplement (Life Technologies), 1% penicillin/streptomycin and 20 ng/mL FGF-2). The emerging rosettes were picked manually, dissociated completely using accutase and plated on a poly-ornithine/laminin-coated plate. NPCs were expanded in N2B27 medium and fed every other day. To differentiate NPCs into neurons, FGF-2 was withdrawn from the N2B27 medium. NPCs and neurons were characterized for stage-specific markers by immunostaining and flow cytometry (NPCs only), expression profile by single-cell RT–PCR and RNA sequencing and electrophysiological property (neurons).

Total RNA extraction. Total RNA of DPCs, iPSCs, NPCs and neurons was extracted using TRIzol reagent (Life Technologies) according to the manufacturer's protocols. Contaminating DNA in RNA samples was removed using TURBO DNase (Life Technologies) according to the manufacturer's protocols. Quality and quantity of DNase-treated RNA were assessed using NanoDrop 1000 (Thermo Scientific).

PCR for exogenous retrovirus DNA silencing. RNA was extracted from iPSCs as previously described using Trizol reagent (Life Technologies). cDNA was generated from the RNA using SuperScript III protocol according to the manufacturer's instructions. PCR was performed using primers listed below at the following cycles: 94 °C for 10 min; 35 repeats of 94 °C for 30 s, 62 °C for 30 s and 72 °C for 1 min; and finally, 72 °C for 7 min. As a positive control, the pMX plasmid of the four vectors used on the reprogramming of the cells was placed along the samples as well as water as a negative template control for amplification. As an additional positive control for the endogenous genes, two hESC lines were used along with our iPSCs: H1 and HUES6 cells. Primers used were as follows. Endo-cMyc: forward, TTG AGG GGC ATC GTC GCG GGA; reverse, GCG TCC TGG GAA GGG AGA TCC. Endo-Klf4: forward, GAA ATT CGC CCG CTC CGA TGA; reverse, CTG TGT GTT TGC GGT AGT GCC. Endo-OCT3/4: forward, TCT TTC CAC CAG GCC CCC GGC TC; reverse, TGC GGG CGG ACA TGG GGA GAT CC. Endo-SOX2: forward, GCC GAG TGG AAA CTT TTG TCG; reverse, GGC AGC GTG TAC TTA TCC TTC T. Exo transgenes pMXs-TgUS: forward, GTG GTG GTA CGG GAA ATC AC. Exo-Oct4 pMXs-Oct3/4-TgDS: reverse, TAG CCA GGT TCG AGA ATC CA. Exo-Sox2 pMXs-Sox2-TgDS: reverse, GGT TCT CCT GGG CCA TCT TA. Exo-Klf4 pMXs-Klf4-TgDS: reverse, GGG AAG TCG CTT CAT GTG AG. Exo-c-Myc pMXs-c-Myc-TgDS: reverse, AGC AGC TCG AAT TTC TTC CA.

Embryoid body formation for pluripotency characterization. Partly dissociated iPSCs were re-suspended in embryoid body medium (DMEM/F12 medium, 1× N2 supplement and 1% FBS) and cultured on shaker (95 r.p.m.) at 37 °C. Medium was changed every 3–4 days. After 20 days, total RNA of embryoid bodies was extracted for further gene expression analyses by qPCR.

Mycoplasma testing. All tissue culture samples were routinely tested for mycoplasma by PCR. One millilitre of media supernatants (with no antibiotics or fungizone) was collected for all cell lines, spun down and resuspended in TE buffer. Ten microlitres of each sample were used in PCR reaction with the following primers: forward, GGC GAA TGG GTG AGT AAC; reverse, CGG ATA ACG CTT GCG ACC T. Any positive sample was immediately discarded.

Microarray. Three hundred nanograms of total extracted RNA from each sample were subjected to microarray by using the Affymatrix GeneChip one-cycle target labelling kit (Affymatrix, Santa Clara, California) according to the manufacturer's recommended protocols. The resultant biotinylated cRNA was fragmented and then hybridized to the GeneChip Human 1.0 ST Array (764,885 probes, 28,869 genes, 19,734 gene-level probe sets with putative full-length transcript support (GenBank and RefSeq)) on the basis of human genome, Hg18. Arrays were prepared at the University of California DNA Core Facility. Arrays were analysed by the Affy (Affymetrix pre-processing)[29] Bioconductor software package for microarray data. Data were then normalized by the RMA (robust multichip averaging) method to background-corrected and normalized probe levels to obtain a summary expression of normalized values for each probe set. Normalized microarray samples were then clustered by a hierarchical approach based on a matrix of distances. Normalized expression data were used to create a distance matrix that was calculated on the basis of Euclidean distance between the transcripts over a pair of samples representing a variation between two samples. Having the distances for all pairs of samples, a linkage method is used to cluster samples in a dendrogram by using calculated distances (sample expression similarities). This method also creates a heat map to graphically show the expression correlation between the samples.

Gene expression analyses by qPCR. RNA samples were reverse transcribed into cDNA using the Super Script III First Strand Synthesis System (Invitrogen, California) according to the manufacturer's instructions. Reactions were run on the Bio-Rad detection system using Sybr-green master mix (Bio-Rad). Primers were selected from Primerbank; validated database (http://pga.mgh.harvard.edu/primerbank/) and specificities were confirmed by melting curve analysis through a Bio-Rad detection system. Sequences of the primers are described in Supplementary Table 12. Quantitative analysis used the comparative threshold cycle method[30]. GAPDH was used as housekeeping gene. Each sample was run in triplicate.

RNA-seq and global gene expression analyses. The RNA-seq analyses were previously described by our group[31]. Briefly, RNAs were isolated using the RNeasy Mini kit (Qiagen). A total of 1,000 ng of RNA was used for library preparation using the Illumina TruSeq RNA Sample Preparation Kit. The RNAs were sequenced on Illumina HiSeq2000 with 50 bp paired-end reads, generating 50 million high-quality sequencing fragments per sample on average. For validation purposes of biological samples subjected to RNA-seq, hESC and iPSC data available from the literature were downloaded and used to compare with our sequenced cell lines. The two hESC lines used are available (HUES-6, referred as ES(HUES), SRR873630, http://www.ncbi.nlm.nih.gov/sra/SRX290739; and H1, referred to here as ES(H1), SRR873631, http://www.ncbi.nlm.nih.gov/sra/SRX290740). The two human iPSC lines used are available under accession codes SRR873619 (referred to here as iPS(TD,1)) and SRR873620 (referred to here as iPS(TD,2)).

Gene ontology (GO) enrichment analysis. RNA-seq enrichment used WebGestalt[32] and Cytoscape[33] software plugins, considering only categories having statistical significance ($P < 0.05$). Genes tested for differential expression were used as the background for GO annotation and enrichment analysis.

NPC counting. NPCs were seeded onto poly-ornithine/laminin-coated six-well plates at a total number of 10^5 cells per well on day 0. Medium change was done on day 2. Cells were collected and counted on day 4.

NPC flow characterization. NPCs were resuspended, dissociated with accutase and fixed using fixation buffer (BioLegend) for 15 min followed by three PBS washes. The cell pellet was incubated and kept in Perm III buffer (BD Biosciences) at −20 °C until needed for the experiment. A total of 10^6 cells were incubated with antibodies Sox1 (PE), Sox2 (APC) or Nestin (PE) and Pax6 (APC) (Bd Biosciences) for 30 min and then washed three times before being resuspended for cell analyses. Cells were analysed in a plate reader mode using FACS Canto II machine (BD Biosciences).

Immunofluorescence staining. Cells were fixed in 4% paraformaldehyde for 10–20 min, washed with PBS three times (5 min each), permeabilized with 0.1% Triton X-100 for 15 min, incubated in blocking solution (2% BSA) for 1 h at room temperature and then in primary antibodies (goat anti-Nanog, Abcam ab77095, 1:500; rabbit anti-Lin28, Abcam ab46020, 1:500; rabbit anti-Oct4, Abcam ab19857, 1:500; mouse anti-SSEA4, Abcam ab16287, 1:200; mouse anti-Nestin, Abcam ab22035, 1:200; rabbit anti-Musashi1, Abcam ab52865, 1:250; rat anti-CTIP2, Abcam ab18465, 1:250; rabbit anti-SATB2, Abcam ab34735, 1:200; chicken anti-MAP2, Abcam ab5392, 1:1,000; rabbit anti-FZD9, Origene TA314730, 1:150; chicken anti-EGFP, Abcam ab13970, 1:1,000; rabbit anti-Synapsin1, EMD-Millipore AB1543P, 1:500; mouse anti-Vglut1, Synaptic Systems 135311, 1:500; rabbit anti-Homer1, Synaptic Systems 160003, 1:500) overnight at 4 °C. The next day, cells were washed with PBS three times (5 min each), incubated with secondary antibodies (Alexa Fluor 488, 555 and 647, Life Technologies, 1:1,000) for 1 h at room temperature and washed with PBS three times (5 min each). Nuclei were stained using DAPI (1:10,000). Slides or coverslips were mounted using ProLong Gold antifade mountant (Life Technologies).

DNA fragmentation analysis. One million NPCs were harvested to single-cell suspension in 1mL PBS, then fixed by addition of 3 mL of 100% ethanol and stored at 4 °C for at least 2 h. NPC pellets were washed once with 5 mL PBS. After removal of PBS, cells were resuspended in 1 mL of propidium iodide (PI) staining solution (0.1% (v/v) Triton X-100, 10 μg/mL PI and 100 μg/mL RNase A in 1× PBS). WS and TD NPC samples were analysed by FACS on a Becton Dickinson LSRI, and gating of subG1 population (cells with fragmented DNA) was examined using FlowJo flow cytometry analysis software.

Caspase assay. Caspase activity was measured using a Green FLICA Caspases 3 & 7 Assay Kit (ImmunoChemistry Technologies). Briefly, NPCs were harvested, washed and stained with 1× carboxyfluorescein Fluorochrome Inhibitor of Caspase Assay (FAM-FLICA) reagent, 10 μg/mL Hoechst and 10 μg/mL propidium iodide (PI). Samples were analysed on the NC-3000 using the pre-optimized Caspase Assay. The population with caspase activity was used to analyse for apoptosis.

Proliferation assay. NPC proliferation was assessed using BD Pharmingen BrdU Flow Kits (BD Biosciences) according to the manufacturer's protocol. Briefly, NPCs were incubated with 1 μM BrdU for 45 min at 37 °C and harvested to single-cell suspension. NPCs were then fixed and permeabilized using BD Cytofix/Cytoperm Buffer and stained using FITC-conjugated anti-BrdU antibody and 7-aminoactinomycin D (7-AAD), a fluorescent dye for labelling DNA. Fluorescence-activated cell sorting (FACS) was done on LSRFortessa (BD Biosciences) and, to obtain the percentage of the BrdU-positive population, the cell-cycle profiles were analysed using FlowJo flow cytometry analysis software.

Construction and characterization of lentiviruses. Commercially available lentiviral vectors (pLKO.1) expressing short hairpin RNAs (shRNAs) against FZD9 under the control of the U6 promoter (Thermo Scientific) were engineered to express the *Discosoma* sp. red fluorescent protein (RFP) mCherry under the control of the hPGK (human phosphoglycerate kinase) promoter. The following shRNAs against *FZD9* and a non-silencing scrambled control shRNA were selected (Thermo Scientific): shRNA-control, 5′-TTC TCC GAA CGT GTC ACG T-3′; shRNA-FZD9, 5′-ATC TTG CGG ATG TGG AAG AGG-3′. For rescue experiments, FZD9 cDNA was amplified from TD NPC cDNA as template by the following primer pair: 5′-CCG AGA TCT TCG AGG TGT GTG GGG TTC TCC AAA G-3′; 5′-TCT AGA GCC ACC ATG GCC GTA GCG CCT CTG-3′.

The reaction was performed using Phusion High-Fidelity DNA polymerase (New England Biolabs) according to the manufacturer's protocol. The FZD9 cDNA was cloned into a lentiviral vector driven by the ubiquitin promoter followed by a self-clevage peptide and GFP sequence. The specificity and efficiency of shRNA-control, shRNA-FZD9, and the FZD9-WT constructs were verified by co-transfection into HEK-293 cells. Cell lysates were collected and analysed by western blot analysis with anti-FZD9 antibodies (Aviva OAEC02415, 1:1,000).

CHIR-98014 treatment. CHIR-98014 (Selleckchem) was resuspended according to manufacturer's instructions into 10 mM stock using DMSO and then diluted to 100 μM. Final concentration used in cells was 100 nM of CHIR-98014, whereas the vehicle cells received only DMSO. For qPCR experiments, NPCs were propagated in six-well plates until 70% confluency and then treated with CHIR-98014 for 6 h to have their RNA collected using Trizol as previously described. For the NPC counting experiment, cells were seeded in six-well plates as described in the presence of CHIR-98014 or DMSO, in triplicates (TD and WS). After 48 h, the culture medium was changed and treatment was repeated. Cells were collected and counted after 96 h of incubation.

Astrocyte differentiation. The TD NPCs were lifted into suspension and maintained on a shaker (95 r.p.m.) to form neurospheres for 3 weeks. For the first week, the spheres were grown with N2B27 medium. The neurospheres were overlaid with the astrocyte medium (Lonza) for the remaining 2 weeks. The neurospheres were plated onto poly-ornithine- and laminin-coated plates and expanded for two to three passages before experimentation. Co-cultures of neurons and astrocytes were prepared for morphometric and functional analyses.

Western blotting. NPCs were lysed in RIPA buffer with protease inhibitor. Rabbit anti-FZD9 antibody (Aviva OAEC02415, 1:1,000) and mouse anti-β-actin (Abcam ab8226, 1:3,000) were used as primary antibodies. IRDye 800CW goat anti-rabbit and IRDye 680RD goat anti-mouse (1:10,000) were used as secondary antibodies. The Odyssey system was used for signal detection. Signal intensities were measured using the Odyssey Image Studio and semi- quantitative analysis of FZD9 signal intensity was corrected with respect to β-actin relative quantification. A paired *t*-test analysis with $P < 0.05$ was used in the comparison of TD and WS FZD9 signal intensity normalized data.

Synaptic puncta quantification. Co-localized Vglut (presynaptic) and Homer1 (postsynaptic) puncta were quantified after three-dimensional reconstruction of *z*-stack random images for all individuals and from two different experiments. Slides were analysed under a fluorescence microscope (Z1 Axio Observer Apotome, Zeiss). Only puncta in proximity of MAP2-positive processes were scored.

Single-cell qRT–PCR and analysis. Specific target amplification was performed in individual dissociated NPCs or 6-week-old neurons using C1 Single-Cell and BioMark HD Systems (Fluidigm), according to the manufacturer's protocol and as described previously[34–36]. Briefly, single cells were captured on a C1 chip (10- to 17-μm cells) and cell viability was checked using a LIVE/DEAD Cell Viability/ Cytotoxicity kit (Life Technologies). After lysis, RNA was reverse transcribed into cDNA with validated amplicon-specific DELTAgene Assays (Supplementary Table 13) using SuperScript III RT Platinum Taq Mix. Specific target amplification was performed by 18 cycles of 95 °C denaturation for 15 s and 60 °C annealing and amplification for 4 min. Each preamplified cDNA was mixed with 2× SsoFast EvaGreen Supermix with Low ROX (Bio-Rad) and then pipetted into an individual sample inlet in a 96.96 Dynamic Array IFC chip (Fluidigm). DELTAgene primer pairs (Supplementary Table 13) were diluted and pipetted into individual assay inlets in the same 96.96 Dynamic Array IFC chip. Quantitative PCR results were analysed using Fluidigm's Real-time PCR Analysis software using the linear (derivative) baseline correction method and the automatic (gene) C_t threshold method with 0.65 curve quality threshold. Hierarchical clustering heat map, PCA analyses, violin plots of $\log_2$(expression of C_t values) (limit of detection = 24) and ANOVA statistical analysis were performed using Singular Analysis Toolset 3.0 (Fluidigm).

Calcium imaging. Neuronal networks derived from human iPSCs were transduced with lentivirus carrying the Syn::RFP reporter construct. Cell cultures were washed with Krebs HEPES buffer (KHB) (10 mM HEPES, 4.2 mM $NaHCO_3$, 10 mM dextrose, 1.18 mM $MgSO_4$, 1.18 mM KH_2PO_4, 4.69 mM KCl, 118 mM NaCl, 1.29 mM $NaCl_2$; pH 7.3) and incubated with 2–5 μM Fluo-4AM (Molecular Probes/ Invitrogen, Carlsbad, California) in KHB for 40 min. Five thousand frames were acquired at 28 Hz with a region of 256 pixels × 256 pixels (×100 magnification), using a Hamamatsu ORCA-ER digital camera (Hamamatsu Photonics K.K., Japan) with a 488 nm (FITC) filter on an Olympus IX81 inverted fluorescence confocal microscope (Olympus Optical, Japan). Images were acquired with MetaMorph 7.7 (MDS Analytical Technologies, Sunnyvale, California), processed and analysed using individual circular regions of interest (ROI) on ImageJ and Matlab 7.2 (Mathworks, Natick, Massachusetts). Syn::RFP^+ neurons were selected after confirmation that calcium transients were blocked with 1 mM of tetrodotoxin (TTX). The amplitude of signals was presented as relative fluorescence changes ($\Delta F/F$) after background subtraction. The threshold for calcium spikes was set at the 95th percentile of the amplitude of all detected events.

Electrophysiology. For whole-cell patch-clamp recordings, individual coverslips containing live 1-month-old neurons were transferred into a heated recording chamber and continuously perfused (1 mL/min) with artificial cerebrospinal fluid bubbled with a mixture of CO_2 (5%) and O_2 (95%) and maintained at 25 °C. Artificial cerebrospinal fluid contained (in mM) 121 NaCl, 4.2 KCl, 1.1 CaCl2, 1 $MgSO_4$, 29 $NaHCO_3$, 0.45 NaH_2PO_4-H_2O, 0.5 Na_2HPO_4 and 20 glucose (all chemicals from Sigma). Whole-cell recordings were performed using a digidata 1440A/ Multiclamp 700B and Clampex 10.3 (Molecular devices). Patch electrodes were filled with internal solutions containing 130 mM K-gluconate, 6 mM KCl, 4 mM NaCl, 10 mM Na-HEPES, 0.2 mM K-EGTA; 0.3 mM GTP, 2 mM Mg-ATP, 0.2 mM cAMP, 10 mM D-glucose, 0.15% biocytin and 0.06% rhodamine. The pH and osmolarity were adjusted for physiological conditions. Data were all corrected for liquid junction potentials, electrode capacitances were compensated on-line in cell-attached mode and a low-pass filter at 2 kHz was used. The access resistance of the cells in our sample was around 37 MΩ with resistance of the patch pipettes 3–5 MΩ. Spontaneous synaptic AMPA events were recorded at the reversal potential of Cl^- and could be reversibly blocked by AMPA receptor antagonist (10 μM NBQX, Sigma). Spontaneous synaptic GABA events were recorded at the reversal potential of Na^+ and could be reversibly blocked with $GABA_A$ receptor antagonist (10 μM SR95531, Sigma).

Multi-electrode array (MEA). Using 12-well MEA plates from Axion Biosystems, we plated the same density of NPCs from TD and WS individuals in triplicate. Each well was seeded with 10,000 NPCs that were induced into neuronal differentiation as previously described. Each well was coated with poly-L-ornithine and laminin before cell seeding. Cells were fed once a week and measurements were taken before the medium was changed. Recordings were performed using a Maestro MEA system and AxIS software (Axion Biosystems), using a band-pass filter with 10 Hz and 2.5 kHz cutoff frequencies. Spike detection was performed using an adaptive threshold set to 5.5 times the standard deviation of the estimated noise on each electrode. Each plate first rested for 5 min in the Maestro, and then 5–10 min of data were recorded to calculate the spike rate per well. MEA analysis was performed using the Axion Biosystems Neural Metrics Tool, wherein electrodes that detected at least five spikes per minute were classified as active electrodes. Bursts were identified in the data recorded from each individual electrode using an adaptive Poisson surprise algorithm. Network bursts were identified for each well, using a non-adaptive algorithm requiring a minimum of ten spikes with a maximum inter-spike interval of 100 ms. Only channels that exhibited bursting activity (more than ten spikes in 5 min interval) were included in this analysis. After measurement, neurons were immunostained to check morphology and density.

Post-mortem brain specimens and cortical sampling. We used six post-mortem brains (two WS and four TD) that were gender-, age- and hemisphere-matched. All brain specimens were harvested within a post-mortem interval of 18–30 h and had been immersed and fixed in 10% formalin for up to 20 years. For the purpose of the present experiments, samples were obtained from anatomically well-identified cortical areas in a consistent manner across specimens. Tissue blocks approximately 5 mm^3 were removed from primary somatosensory cortex (Brodmann area 3) and primary motor cortex (Brodmann area 4) in the arm/hand knob region of the pre- and postcentral gyri, respectively, and from the secondary visual area (Brodmann area 18) from approximately 1.4 cm dorsally to the occipital pole and 2 cm from the midline[37,38]. We focused specifically on these parts of the cortex because pathologies in dendritic morphology in these areas have been reported in other neurodevelopmental disorders[39–41]. In addition, pyramidal neurons in the selected areas reach their mature-like morphology early in development and start displaying dendritic pathologies sooner than high integration areas, such as the prefrontal cortex, allowing comparison of post-mortem findings with iPSC-derived neurons in early stages of development[42,43].

Post-mortem brain tissue processing. Sampled tissue blocks were processed using an adaptation of the Golgi–Kopsch method[44], which has been shown to give good results with tissue that has been fixed for long periods[45]. Briefly, blocks were immersed in a solution of 3% potassium dichromate, 0.5% formalin for 8 days, followed by immersion into 0.75% silver nitrate for 2 days. Blocks were then sectioned on a vibratome, perpendicular to the pial surface, at a thickness of 120 μm. Golgi sections were cut into 100% ethyl alcohol and transferred briefly into methyl salicylate followed by toluene, mounted onto glass slides and coverslipped. Adjacent blocks from each region were sectioned at 60 μm and stained with thionin for visualization of cell bodies and laminar organization, which enabled identification of the position of each neuron within a specific cortical layer. Cytoarchitectonic analysis of histological sections from each block confirmed that tissue was sampled from the ROI and that the Golgi-impregnated pyramidal neurons were located in cortical layers V/VI.

Morphometric analysis of Golgi-impregnated neurons. Cortical neurons from all six post-mortem brains were used in the study. Neurons included in the morphological analysis did not display degenerative changes[46]. Only neurons with fully impregnated soma, apical dendrites with present oblique branches and at least two basal dendrites with third-order segments were chosen for the analysis[47]. To minimize the effects of cutting on dendritic measurements, we included neurons with cell bodies located near the centre of 120-μm-thick histological sections, with natural terminations of higher-order dendritic branches present where possible[37,47]. Inclusion of the neurons completely contained within 120-μm sections biases the sample towards smaller neurons, leading to the underestimation of dendritic length[48]; therefore, we applied the same criteria blinded across all WS and TD specimens, and we thus included the neurons with incomplete endings if they were judged to otherwise fulfil the criteria for successful Golgi impregnation. All neurons were oriented with apical dendrite perpendicular to the pial surface; inverted pyramidal cells as well as magnopyramidal neurons were excluded from the analysis. Neuronal morphology was quantified along *x*-, *y*-, and *z*-coordinates using Neurolucida version 10 software (MBF Bioscience, Williston, Vermont) connected to a Nikon Eclipse 80i microscope, with a ×40 (0.75 numerical aperture) Plan Fluor dry objective. Tracings were conducted on both apical and basal dendrites, and the results reflect summed values for both types of dendrite per neuron. Following the recommendation that the applications of Sholl's concentric spheres or Eayrs' concentric circles for the analysis of neuronal morphology are not adequate when neuronal morphology is analysed in three dimensions[48], we conducted dendritic tree analysis with the following measurements[37,47]: (1) soma area—cross sectional surface area of the cell body; (2) dendritic length—summed total length of all dendrites per neuron; (3) dendrite number—number of dendritic trees emerging directly from the soma per neuron; (4) dendritic segment number—total number of segments per neuron; (5) dendritic spine/protrusion number—total number of dendritic spines per neuron; (6) dendritic spine/protrusion density—average number of spines per 20 μm of dendritic length; and (7) branching point number—number of nodes (points at the dendrite where a dendrite branches into two or more) per neuron. Dendritic segments were defined as parts of the dendrites between two branching points—between the soma and the first branching point in the case of first-order dendritic segments, and between the last branching point and the termination of the dendrite in the case of terminal dendritic segments. Since the long formalin-fixation time could have resulted in degradation of dendritic spines, spine values might be underestimated and are thus reported here with caution. All of the tracings were accomplished blind to brain region and diagnostic status.

Morphometric analysis of iPSC-derived neurons. The iPSC-derived sample consisted of EGFP-positive 8-week-old neurons with pyramidal- or ovoid-shaped soma and at least two branched neurites (dendrites) with visible spines/protrusions. Protrusions from dendritic shaft, which morphologically resembled dendritic spines in post-mortem specimens, were considered and quantified as dendritic spines in iPSC-derived neurons. The neurites were considered dendrites on the basis of the criteria applied in post-mortem studies: (1) thickness that decreased with the distance from the cell body; (2) branches emerging under acute angle; and (3) presence of dendritic spines. In addition, only enhanced-GFP-positive neurons with nuclei co-stained with CTIP2, indicative of layer V/VI neurons, and with the dendrites displaying evenly distributed fluorescent stain along their entire length, were included in the analysis. The morphology of the neurons was quantified along *x*-, *y*-, and *z*-coordinates using Neurolucida version 9 software (MBF Bioscience, Williston, VT) connected to a Nikon Eclipse E600 microscope with a ×40 oil objective. No distinction was made between apical and basal dendrites, and the results reflect summed length values of all neurites/dendrites per neuron, consistent with what was done for the post-mortem neurons. The same set of measurements used in the analysis of Golgi-impregnated neurons was applied to the analysis of iPSC-derived neurons, and all of the tracings were accomplished blind to the diagnostic status and were conducted by the same rater (B.H.-M.). Intra-rater reliability was assessed by having the rater trace the same neuron after a period of time. The average coefficient of variation between the results of retraced neurons was 2% for soma area (SA), total dendritic length (TDL), dendritic segment number (DSN) and branching point number (BPN), and 3% for dendritic spine/protrusion number (DPN); there was no variation in tree/dendrite number (TN) in different tracings of the same neuron. The accuracy was further checked by having three individuals (B.H.-M., B.J. and L.S.) trace the same neuron.

Brain imaging data acquisition and quality control. MRI scanning was completed in 19 participants with WS (aged 19–43 years; mean 29.0, s.d. 8.8; 11 males, 8 females) and 19 TD comparison participants (aged 16–43 years; mean 26.2, s.d. 7.3; 8 males, 11 females). There was no significant difference between the groups in age ($t = 1.0$, $P < 0.30$) or in gender ratio (Pearson's $\chi^2 = 0.95$, $P < 0.33$). A standardized multiple modality high-resolution structural MRI protocol was implemented, involving three-dimensional T_1- and T_2-weighted volumes and a set of diffusion-weighted scans. Imaging data were obtained at the University of California San Diego Radiology Imaging Laboratory on a 1.5 T GE Signa HDx 14.0M5 TwinSpeed system (GE Healthcare, Waukesha, Wisconsin) using an eight-channel phased array head coil. A three-dimensional inversion recovery spoiled gradient echo (IR-SPGR) T_1-weighted volume was acquired with pulse sequence parameters optimized for maximum grey/white matter contrast (echo time = 3.9 ms, repetition time = 8.7 ms, inversion time = 270 ms, flip angle = 8°, difference in echo times = 750 ms, bandwidth = ± 15.63 kHz, field of view = 24 cm, matrix = 192 × 192, voxel size = 1.25 mm × 1.25 mm × 1.2 mm). All MRI data were collected using prospective motion (PROMO) correction for non-diffusion imaging[49]. This method has been shown to improve image quality, reduce motion-related artefacts, increase the reliability of quantitative measures and improve the clinical diagnostic utility of MRI data obtained in children and clinical groups[50,51]. Standardized quality control procedures were followed for both raw and processed data, including visual inspection ratings by a trained imaging technician and computer algorithms testing general image characteristics as well as aspects specific to each imaging modality, such as contrast properties, registrations and artefacts from motion and other sources. Participants included in the current analyses were only those who passed all raw and processed quality control measures.

MRI data post-processing. Image post-processing and analysis were performed using FreeSurfer software suite (http://surfer.nmr.mgh.harvard.edu/). Surface-based cortical reconstruction and subcortical volumetric segmentation procedures have been shown elsewhere[52–58]. Briefly, a three-dimensional model of the cortical surface was generated using MRI scans with four attributes: white matter segmentation; tessellation of the grey/white matter boundary; inflation of the folded, tessellated surface; and correction of topological defects[53,54]. Cortical thickness was measured using the distances from each point on the white matter surface to the pial surface[57]. Cortical surface area was measured at the pial surface for the entire cerebrum and for each parcel of the Desikan and Destrieux atlases[53,54,58,59].

Statistical analysis. Means ± s.e.m. for each parameter were obtained from samples described in Supplementary Table 1. There were no statistical methods used to predetermine sample size. The experiments were not randomized. All of the tracings were accomplished blind to brain region and diagnostic status. All statistical analyses were done using Prism (Graphpad). Before statistical analysis comparing means between three to five unmatched groups of data, normal distribution was tested using D'Agostino and Pearson omnibus normality test and variance similarity was tested using Bartlett's test for equal variances. Means of three to five unmatched groups, where normal distribution and equal variances between groups were confirmed, were statistically compared using one-way ANOVA and Tukey's post hoc test. Otherwise, a Kruskal–Wallis test and Dunn's multiple comparison test were used. Before statistical analysis comparing means between two unmatched groups of data, normal distribution was tested using D'Agostino and Pearson omnibus normality test and variance similarity was tested using an *F* test to compare variances. To compare the means of two groups where normal distribution and similar variance between groups were confirmed, Student's *t* test was used. Otherwise, a Mann–Whitney test was used. Significance was defined as $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$ or $^{****}P < 0.0001$.

26. Lawless, J. F. & Fredette, M. Frequentist prediction intervals and predictive distributions. *Biometrika* **92,** 529–542 (2005).
27. Chen, J. *et al.* Transcriptome comparison of human neurons generated using induced pluripotent stem cells derived from dental pulp and skin fibroblasts. *PLoS ONE* **8,** e75682 (2013).
28. Marinho, P. A., Chailangkarn, T. & Muotri, A. R. Systematic optimization of human pluripotent stem cells media using Design of Experiments. *Sci. Rep.* **5,** 9834 (2015).
29. Gautier, L., Cope, L., Bolstad, B. M. & Irizarry, R. A. affy—analysis of *Affymetrix GeneChip* data at the probe level. *Bioinformatics* **20,** 307–315 (2004).
30. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods* **25,** 402–408 (2001).
31. Marchetto, M. C. *et al.* Differential L1 regulation in pluripotent stem cells of humans and apes. *Nature* **503,** 525–529 (2013).
32. Zhang, B., Kirov, S. & Snoddy, J. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.* **33,** W741–W748 (2005).
33. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
34. Llorens-Bobadilla, E. *et al.* Single-cell transcriptomics reveals a population of dormant neural stem cells that become activated upon brain injury. *Cell Stem Cell* **17,** 329–340 (2015).
35. Livak, K. J. *et al.* Methods for qPCR gene expression profiling applied to 1440 lymphoblastoid single cells. *Methods* **59,** 71–79 (2013).
36. Hermann, B. P. *et al.* Transcriptional and translational heterogeneity among neonatal mouse spermatogonia. *Biol. Reprod.* **92,** 54 (2015).
37. Jacobs, B. *et al.* Regional dendritic and spine variation in human cerebral cortex: a quantitative golgi study. *Cereb. Cortex* **11,** 558–571 (2001).
38. Semendeferi, K. *et al.* Spatial organization of neurons in the frontal pole sets humans apart from great apes. *Cereb. Cortex* **21,** 1485–1497 (2011).
39. Marin-Padilla, M. Structural abnormalities of the cerebral cortex in human chromosomal aberrations: a Golgi study. *Brain Res.* **44,** 625–629 (1972).
40. Takashima, S., Becker, L. E., Armstrong, D. L. & Chan, F. Abnormal neuronal development in the visual cortex of the human fetus and infant with down's syndrome. A quantitative and qualitative Golgi study. *Brain Res.* **225,** 1–21 (1981).
41. Jay, V., Chan, F. W. & Becker, L. E. Dendritic arborization in the human fetus and infant with the trisomy 18 syndrome. *Brain Res. Dev. Brain Res.* **54,** 291–294 (1990).
42. Marin-Padilla, M. Prenatal and early postnatal ontogenesis of the human motor cortex: a golgi study. II. The basket-pyramidal system. *Brain Res.* **23,** 185–191 (1970).
43. Vukšić, M., Petanjek, Z., Rasin, M. R. & Kostović, I. Perinatal growth of prefrontal layer III pyramids in Down syndrome. *Pediatr. Neurol.* **27,** 36–38 (2002).
44. Jacobs, B. *et al.* Quantitative analysis of cortical pyramidal neurons after corpus callosotomy. *Ann. Neurol.* **54,** 126–130 (2003).
45. Riley, J. N. A reliable Golgi-Kopsch modification. *Brain Res. Bull.* **4,** 127–129 (1979).
46. Williams, R. S., Ferrante, R. J. & Caviness, V. S., Jr. The Golgi rapid method in clinical neuropathology: the morphologic consequences of suboptimal fixation. *J. Neuropathol. Exp. Neurol.* **37,** 13–33 (1978).
47. Jacobs, B. & Scheibel, A. B. A quantitative dendritic analysis of Wernicke's area in humans. I. Lifespan changes. *J. Comp. Neurol.* **327,** 83–96 (1993).
48. Uylings, H. B., Ruiz-Marcos, A. & van Pelt, J. The metric analysis of three-dimensional dendritic tree patterns: a methodological review. *J. Neurosci. Methods* **18,** 127–151 (1986).
49. White, N. *et al.* PROMO: Real-time prospective motion correction in MRI using image-based tracking. *Magn. Reson. Med.* **63,** 91–105 (2010).
50. Brown, T. T. *et al.* Prospective motion correction of high-resolution magnetic resonance imaging data in children. *Neuroimage* **53,** 139–145 (2010).
51. Kuperman, J. M. *et al.* Prospective motion correction improves diagnostic utility of pediatric MRI scans. *Pediatr. Radiol.* **41,** 1578–1582 (2011).

52. Jovicich, J. *et al.* Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. *Neuroimage* **30,** 436–443 (2006).
53. Dale, A. M., Fischl, B. & Sereno, M. I. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* **9,** 179–194 (1999).
54. Fischl, B., Sereno, M. I. & Dale, A. M. Cortical surface-based analysis. II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9,** 195–207 (1999).
55. Fischl, B. *et al.* Sequence-independent segmentation of magnetic resonance images. *Neuroimage* **23** (Suppl 1), S69–S84 (2004).
56. Fischl, B. *et al.* Automatically parcellating the human cerebral cortex. *Cereb. Cortex* **14,** 11–22 (2004).
57. Fischl, B. & Dale, A. M. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proc. Natl Acad. Sci. USA* **97,** 11050–11055 (2000).
58. Desikan, R. S. *et al.* An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31,** 968–980 (2006).
59. Destrieux, C., Fischl, B., Dale, A. & Halgren, E. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* **53,** 1–15 (2010).

a

Subject Identifier (**WS**)	**WS25**	**WS79**	**WS77**	**WS17**	**pWS88**
Age	13.64	13.56	8.14	9.73	14.79
Gender	Female	Male	Female	Male	Male
WS Total Score*	9	12	13	14	7
Facial Features; 8-17 items† score 3 points	3	3	3	3	0
Cardiovascular; 1-2 items‡ score 5 points	5	5	5	5	5
Other Cardiovascular; 1-3 items§ score 1 point	1	1	1	1	1

* DSS Total >/=3 indicates WS a possibility. WS mean = 9, SD = 2.2

† e.g. wide mouth; long philtrum; bitemporal narrowing; small, widely spaced teeth prominent ear lobes and lips; broad brow; malocclusion; stellate lacy iris; full cheeks

‡ *SVAS*; peripheral pulmonary artery stenosis

§ Other congenital heart disease; heart murmur; hypertension

Subject Identifier (**Control**)	**TD55**	**TD59**	**TD63**	**TD70**
Age	8.10	9.83	19.72	9.25
Gender	Female	Male	Female	Male

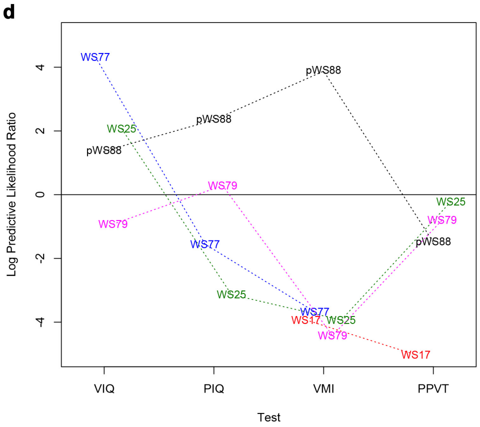
b

Gene Copy number determined by qPCR	WS17	WS25	WS77	WS79	pWS88
CALN-ANY	2	2	2	2	2
NSUN5ex2-1-487T	1	1	1	1	2
NSUN5ex8-5-767A	1	1	1	1	2
TRIM50ex7-ANY	1	1	1	1	2
FKBP6-ANY	1	1	1	1	2
FZD-ANY	1	1	1	1	2
BAZ1B-3-ANY	1	1	1	1	2
BCL7B-ANY	1	1	1	1	2
WBSCR22	1	1	1	1	2
ELN-ANY	1	1	1	1	1
LIMK1-ANY	1	1	1	1	1
EIF4H-ANY	1	1	1	1	1
LAT2-ANY	1	1	1	1	1
RFC2-ANY	1	1	1	1	1
CLIP2-ANY	1	1	1	1	2
BEN-ex25-ANY	1	1	1	1	2
GTF2Iex7	1	1	1	1	2
GTF2Iex12	1	1	1	1	2
GTF2Iex18	2	2	2	2	2
WBSCR16-ANY	2	2	2	2	2

c

	Test	Reference(s)*	Purpose
1. Neurocognitive			
Intelligence	Wechsler Abbreviated Scale of Intelligence (WASI)	71	Verbal, performance, and full-scale IQ (VIQ and PIQ)
Non-verbal ability	Judgment of Line Orientation Test (JLO)	72	Visuospatial perception
	Beery-Buktenica Developmental Test of Visual-Motor Integration (VMI)	73	Visuomotor integration
	Benton Facial Recognition Test (BFRT)	74	Unfamiliar face recognition
Verbal ability	Peabody Picture Vocabulary Test (PPVT-III)	75	Receptive vocabulary
2. Social behavior			
Parental questionnaires of social functioning	Salk Institute Sociability Questionnaire (SISQ)	62, 66, 76, 77	Approachability, social-emotional behavior
Experimental measures	Naturalistic use of language - from narrative tasks	62, 78	Assessment of structural language capacities versus the use of language for social and affective purposes

*included in Supplementary references



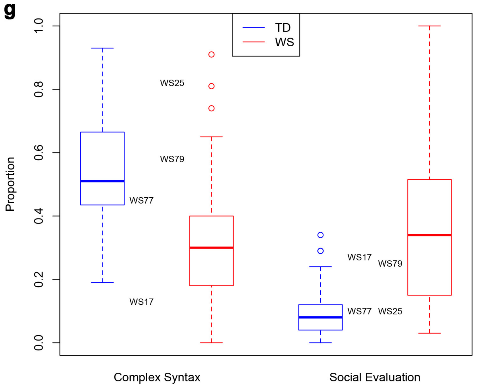
e

Test	N (females)	Age Range (mean)	Mean Standard Score ± SD	Standard Score Range
VIQ	81 (49)	7.0 - 19.9 (14.2)	67.1±10.6	44 - 85
PPVT	97 (58)	7.9 - 19.9 (14.8)	73.5± 15.4	39 - 106
PIQ	81 (49)	7.0 - 19.9 (14.2)	58.3± 10.0	44 - 83
VMI	56 (30)	9.1 - 19.5 (14.2)	52.6± 7.8	44 – 73

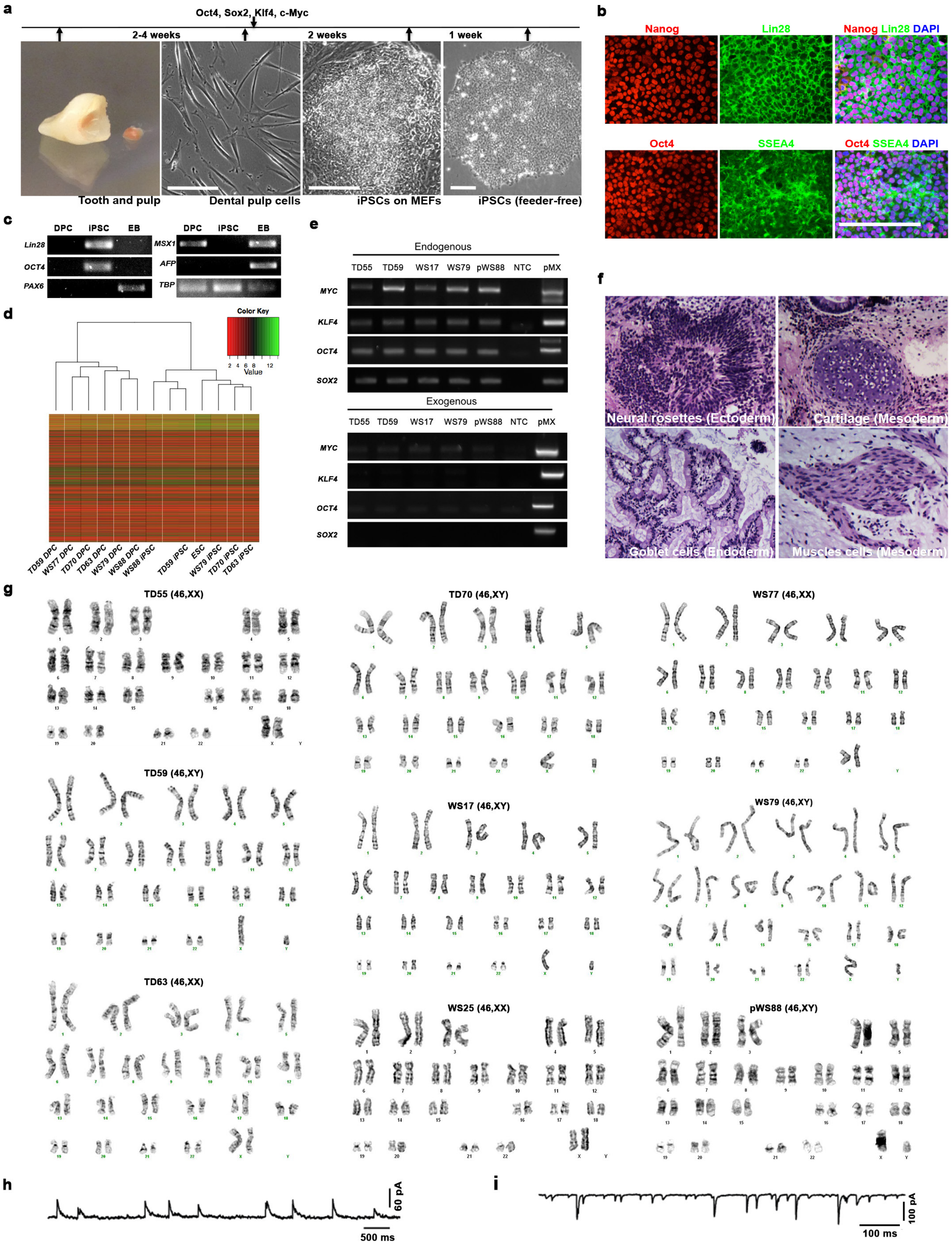
f

	WS	TD
N (females)	65 (38)	22 (16)
Age Range (mean)	10.0 – 19.8 (18.8)	12.5 – 19.96 (18.8)
BFRT Mean ± SD	41.9 ± 5.7	42.9 ± 4.3
BFRT range	25 – 52	32 - 49
JLO Mean ± SD	7.1 ± 5.2	26.4 ± 3.7
JLO Range	0 - 22	14 - 30

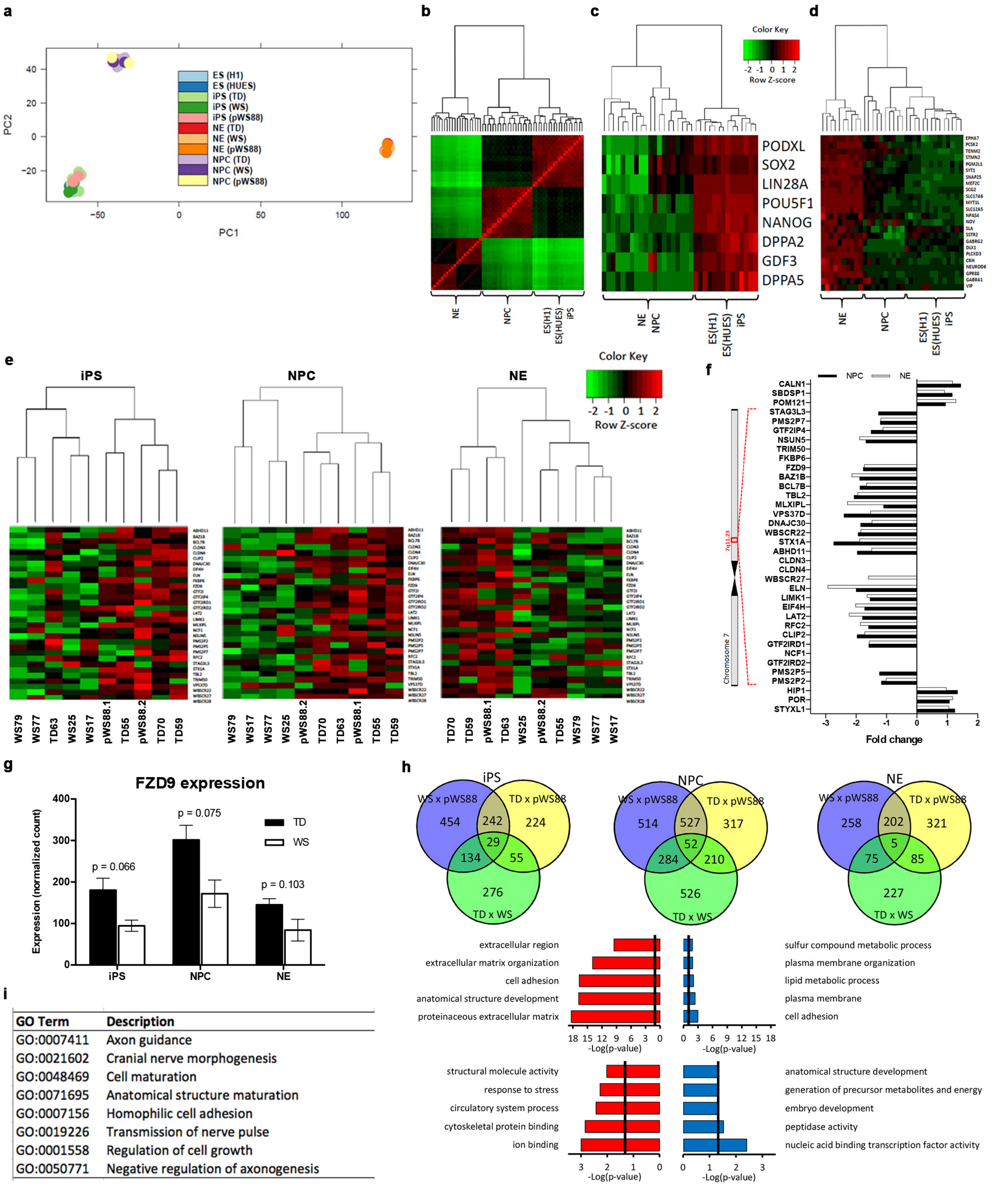
BFRT: Benton Facial Recognition Task
JLO: Judgment of Line Orientation



Extended Data Figure 1 | Participants with WS in iPSC study and their neurocognitive and social profiles. a, Summary of scores on the Diagnostic Score Sheet (DSS) for individuals with WS. **b**, Table showing allele number of genes in WS-deleted region in each participant obtained from qPCR. **c**, Summary of all neurocognitive and social behavioural tests used on this study. **d**, **e**, WS neurocognitive profiles. Log of predictive likelihood ratio for iPSC participants (identified by participant number) calculated as the log of the ratio of the likelihoods for each individual test score based on the predictive distributions for TD individuals and those with WS (**d**). Values less than 0 indicate depressed scores consistent with expectations for WS. Predictive distributions for TD participants used published norms (means and standard deviations with assumed normality). Predictive distributions for individuals with WS were calculated using available WS data (VIQ/PIQ $n = 81$, VMI $n = 56$, PPVT $n = 97$) (**e**), assuming normality and least squares estimation, and according to the procedures described elsewhere[26]. WS parameter estimates for the VMI were calculated using censored regression owing to several individuals with WS scoring at the instrument floor. **f**, Description of population included in Benton Face Recognition and Judgment of Line Orientation in Fig. 1b (TD $n = 22$ versus WS $n = 65$). **g**, Boxplots for WS (red) and TD (blue) participants on complex syntax (WS $n = 45$; TD $n = 47$) and social evaluation (WS $n = 44$; TD $n = 49$). Red and blue circles depict scores more than 1.5 times the interquartile range away from the median.



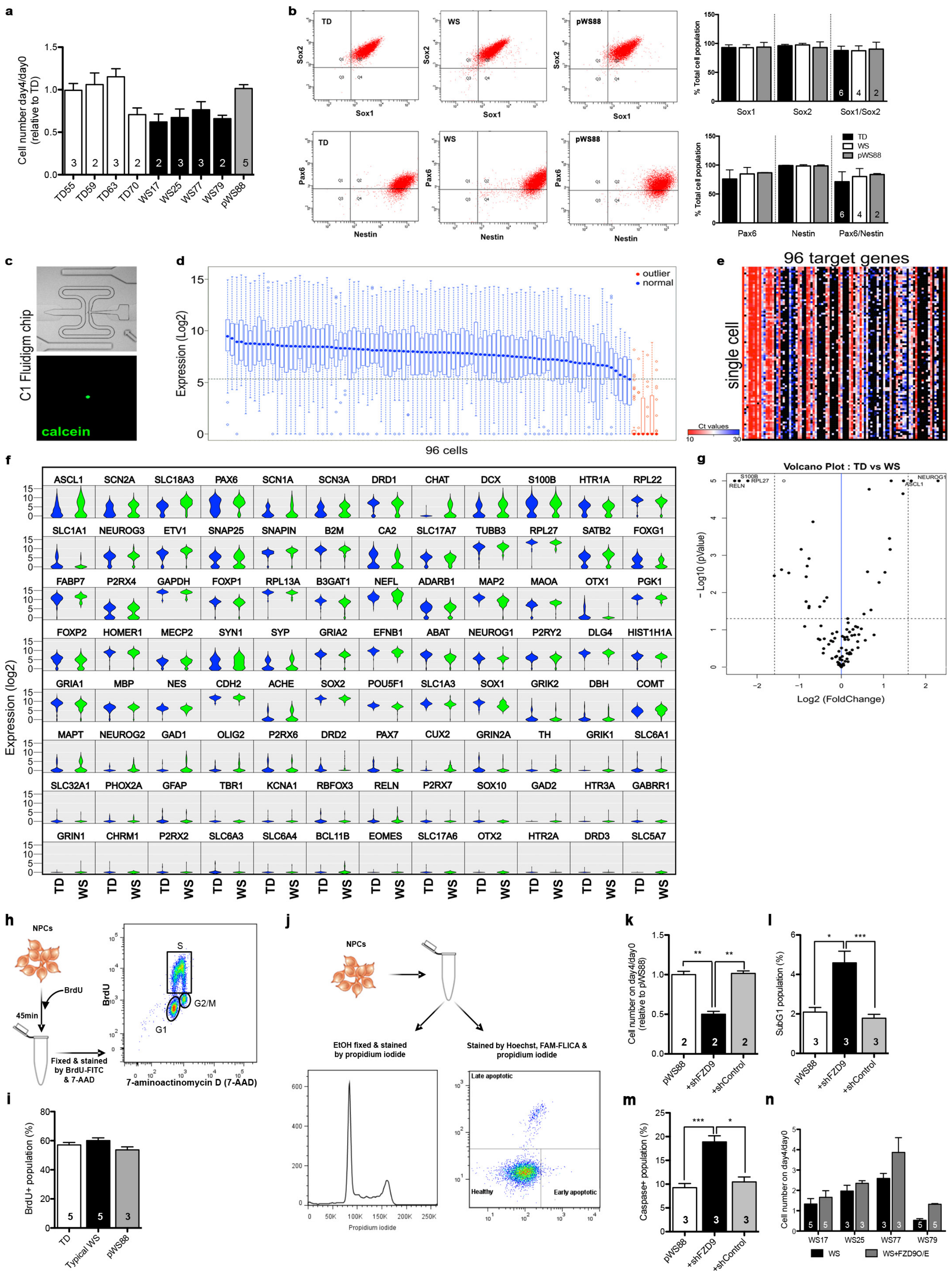
Extended Data Figure 2 | Generation and characterization of iPSCs. **a**, Summary of reprogramming protocol using retrovirus carrying Yamanaka transcription factors (see Supplementary Information for details). Scale bar, 200 μm. **b**, Representative images of iPSCs expressing pluripotent markers including Nanog, Lin28, Oct4 and SSEA4 assessed by immunofluorescence staining. Scale bar, 200 μm. **c**, Expression of three germ-layer markers in iPSC-derived embryoid bodies (EBs); *PAX6* (ectoderm), *MSX1* (mesoderm) and *AFP* (endoderm) assessed by semiquantitative RT–PCR. *TBP*, housekeeping control. **d**, Cluster analysis showing correlation coefficients of microarray profiles of three WS DPCs, three TD DPCs, three WS iPSCs, three TD iPSCs and one ESC. **e**, Representative PCR showing silencing of the four transgenes (exogenous) in iPSCs. **f**, Representative images of teratoma from iPSCs showing tissues of three germ layers; neural rosettes (ectoderm), cartilage (mesoderm), muscle cells (mesoderm) and goblet cells (endoderm). **g**, Representative image of iPSC chromosomes showing its genetic stability assessed by G-banding karyotype analysis. **h**, **i**, Spontaneous synaptic GABA events (**h**) and spontaneous synaptic AMPA events (**i**) in 1-month-old iPSC-derived neurons.



GO Term	Description
GO:0007411	Axon guidance
GO:0021602	Cranial nerve morphogenesis
GO:0048469	Cell maturation
GO:0071695	Anatomical structure maturation
GO:0007156	Homophilic cell adhesion
GO:0019226	Transmission of nerve pulse
GO:0001558	Regulation of cell growth
GO:0050771	Negative regulation of axonogenesis

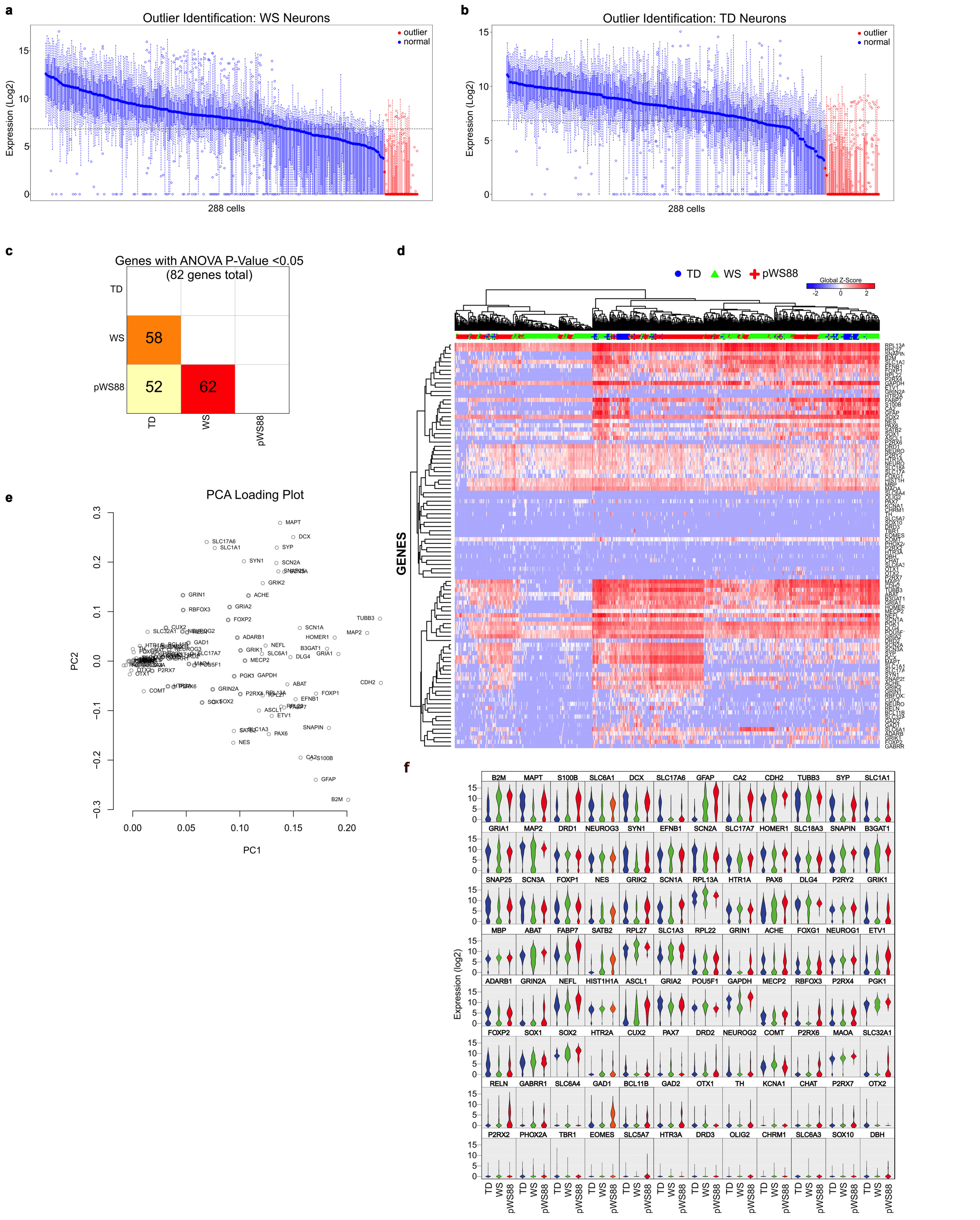
Extended Data Figure 3 | See next page for caption.

Extended Data Figure 3 | Global gene expression analysis during neuronal differentiation. **a**, PCA plot of embryonic stem cells (ES), induced pluripotent stem cells (iPS), neuronal progenitor cells (NPC) and neurons (NE) for TD, WS and pWS88. **c**, Euclidian matrix distance-based heat map and hierarchical clustering-based dendrogram of ES, NPC and NE cells for WD, WS and pWS88 samples. Expression variability between samples is indicated by *Z*-score, varying from green (negative variation) to red (positive variation). **c**, Euclidian matrix distance-based heat map and hierarchical clustering-based dendrogram of pluripotency gene markers for ES, NPC and NE cells for TD, WS and pWS88 samples. **d**, Euclidian matrix distance-based heat map and hierarchical clustering-based dendrogram of neuronal gene markers for iPS, NPC and NE cells for TD, WS and pWS88 samples. Expression variability between samples is indicated by *Z*-score, varying from green (negative variation) to red (positive variation). **e**, Specific cell type-based clustering analysis of biological replicates subjected to RNA-seq for the WS-related genes in three stages during differentiation (iPS, NPC and NE). **f**, Fold change variation of WS-related genes in different cell lines. Ideogram of chromosome 7 (band 7q11.23) corresponding to the commonly deleted region with the WS-related genes. Fold change variation of normalized WS-related gene expression in NPCs and neurons (NE) compared with TDs. Non-represented fold change corresponds to those genes having high expression variability between biological replicates, or having very low expression values. **g**, Expression of *FZD9* gene in iPSC, NPCs and neurons from TD and WS. Error bars, s.e.m. **h**, Venn diagram showing correlation of significant differentially expressed genes between TD, pWS88 and WS during neuronal differentiation. Significantly enriched GO terms found for downregulated (red histogram) and upregulated (blue histogram) differentially expressed genes between TD and WS in NPC. Significantly enriched GO terms found for downregulated (red histogram) and upregulated (blue histogram) differentially expressed genes between TD and WS in neurons (NE). Vertical line (black) corresponds to a significant *P* value (<0.05). **i**, Enriched GO metabolic process terms found in NPC of WS samples correlated with the GO found by a similar comparison performed in ref. 13.

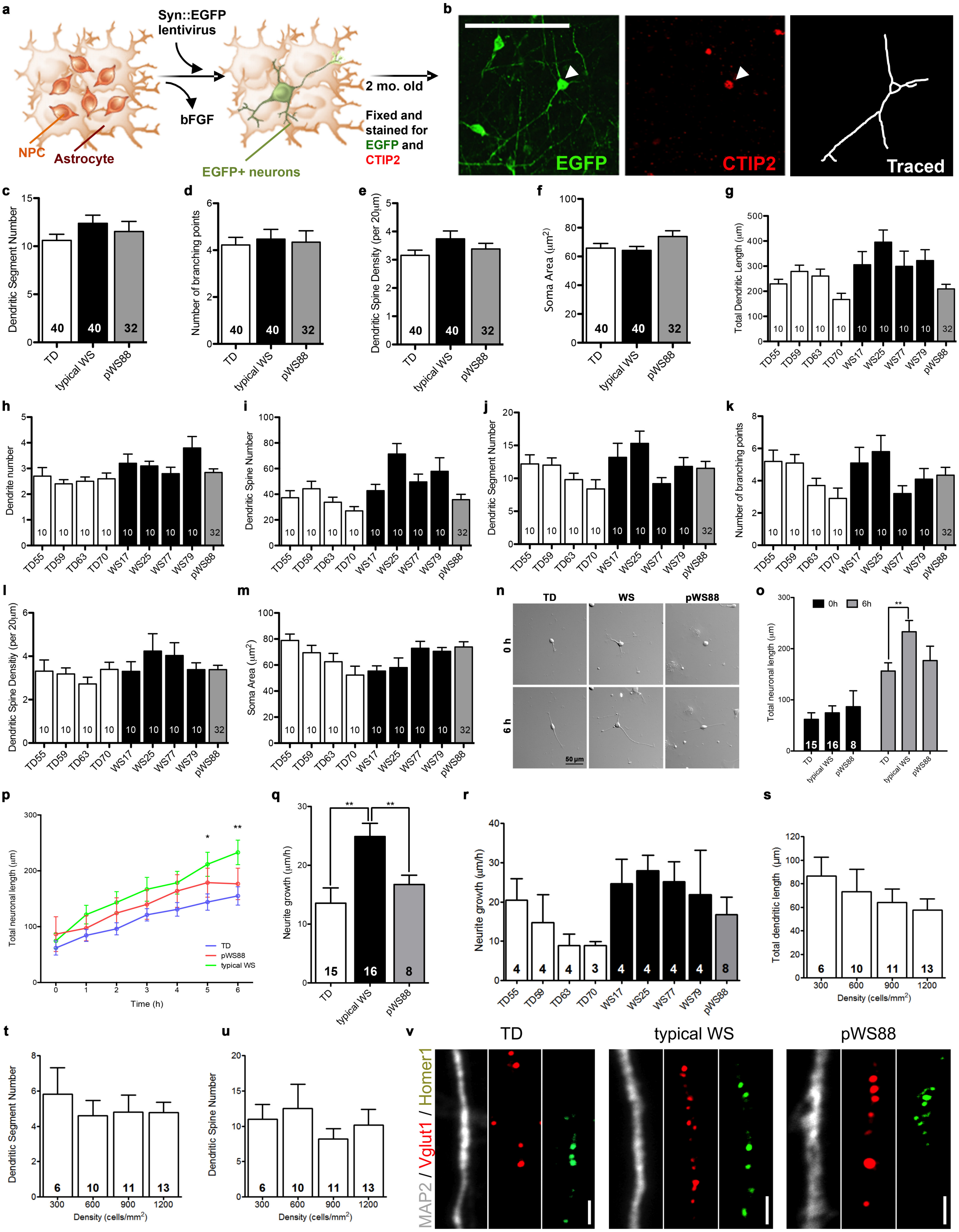


Extended Data Figure 4 | See next page for caption.

Extended Data Figure 4 | Defect in WS NPC apoptosis and role of FZD9. a, Ratio of NPC number on day 4 over day 0 relative to TD. Data are shown as mean ± s.e.m.; *n*, number of clones. **b**, High percentage (>95%) of Sox1/Sox2-positive and Pax6/Nestin-positive cell population was comparably observed in TD, typical WS and pWS88 NPCs assessed by FACS. Data are shown as mean ± s.e.m.; *n*, number of clones. **c**, Microfluidics of C1 chip used to capture live single cells (calcein$^+$ cell). **d**, Outlier exclusion based on the recommended/default limit of detection value of 24, analysed by Fluidigm Singular 3.0. Outliers were removed manually on the basis of the sample median $\log_2$(expression) values. **e**, Representative example of non-normalized C_t plot, indicated with the rectangle in the heat map. Cells are shown in rows and genes in columns. The range of cycle threshold (C_t) values is colour coded from low (blue) to high (red) and absent (black). **f**, Violin plots of all 96 genes showing the comparison between TD and WS NPCs from the single-cell analyses ($\log_2$(expression) values). The majority of genes show unimodal expression distribution. **g**, Volcano plot of single-cell expression data. Plot illustrates differences in expression patterns of target genes of iPSC-derived NPCs. The dotted lines represent more than or equal to 3.0-fold differentially expressed genes between the groups at $P < 0.05$ (unpaired two-sample *t*-test). **h**, Schematic diagram summarizing NPC preparation for proliferation assay and representative scatter plot showing cells in each cycle phase (G1, S and G2/M). **i**, No significant differences in percentage of the BrdU-positive population between TD, typical WS and pWS88 NPCs. **j**, Schematic diagram summarizing NPC preparation for apoptosis analysis and representative analysed data for DNA fragmentation (left) and caspase assay (right). **k**–**m**, Changes in ratio of NPC number on day 4 over day 0 relative to TD (**k**), percentage of subG1 population (**l**) and percentage of population with high caspase activity (**m**) of pWS88 NPCs when treated with shFZD9 and shControl. **n**, Increase in cell number day 4/day 0 upon overexpression of FZD9 in WS iPSC-derived NPCs. Data are shown as mean ± s.e.m. for each individual; *n*, technical replicates. For **i** and **k**–**m**, data are shown as mean ± s.e.m.; *n*, number of clones, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, one-way ANOVA and Tukey's post hoc test (**i**), Kruskal–Wallis test and Dunn's multiple comparison test (**k**–**m**).

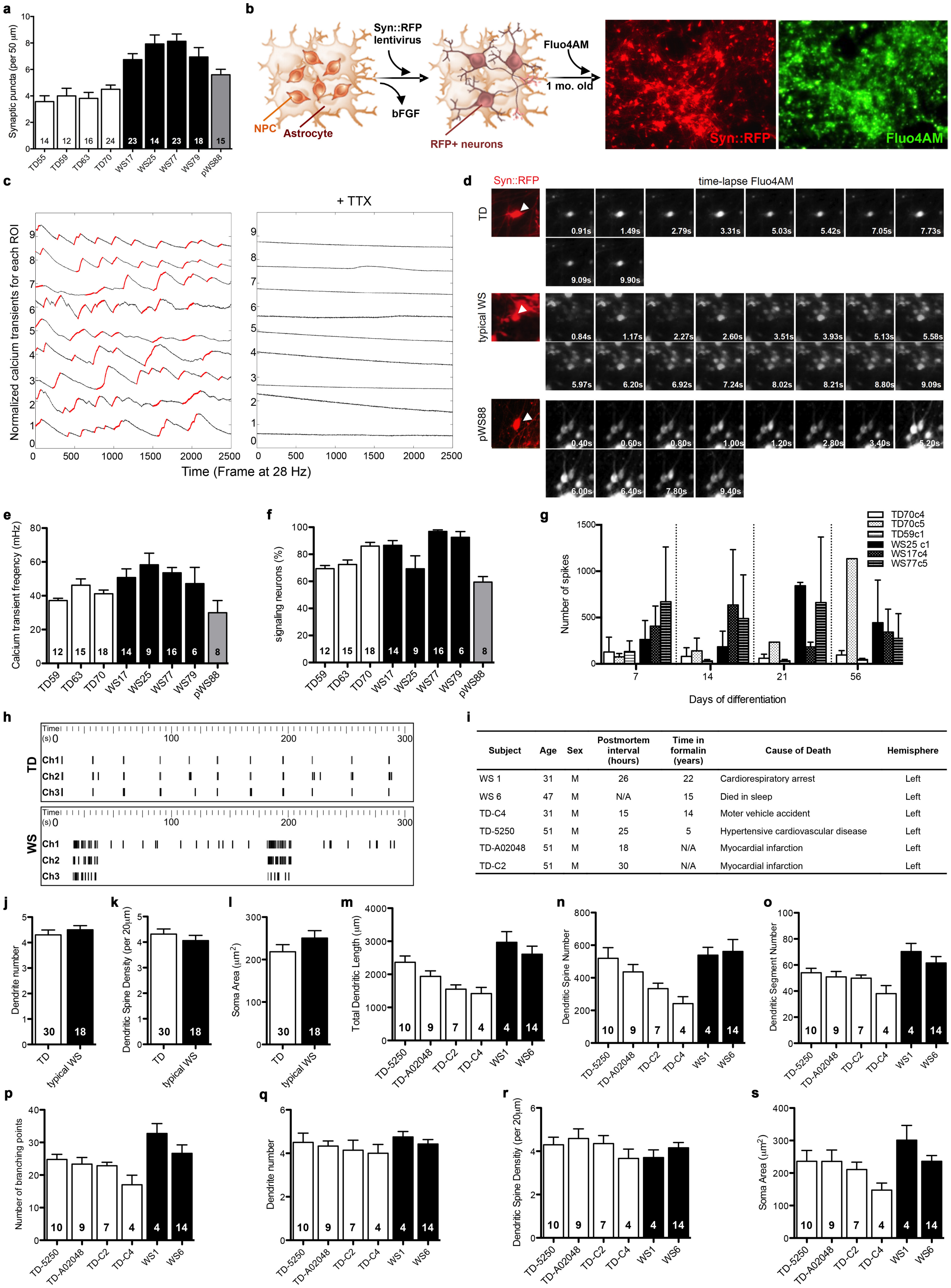


Extended Data Figure 5 | Single-cell analysis of WS and TD iPSC-derived neurons. a, **b**, Outlier exclusion based on limit of detection = 24, analysed by Fluidigm Singular 3.0. Outliers were removed manually on the basis of the sample median $\log_2$(expression) values. **c**, Heat map of number of genes with ANOVA $P < 0.05$ (82 genes in total). **d**, Unsupervised hierarchical clustering of 672 single-cell of WS and TD iPSC-derived neurons identified cell sub-populations not linked with the genotype. Cells are shown in rows and genes in columns. Log_2(gene expression levels) were converted to a global *Z*-score (blue is the lowest value, red is highest). Genes were clustered using the Pearson correlation method and cells were clustered using the Euclidean method. **e**, PCA projections of the 96 genes, showing the contribution of each gene to the first two PCs. **f**, Violin plots of all 96 genes showing the comparison between TD, WS and pWS88 neurons from the single-cell analyses ($\log_2$(expression) values).



Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Morphometric analysis of WS-derived CTIP2-positive cortical neurons. a, Summary of preparation of neurons for evaluation by morphometric analysis. **b**, Representative images of EGFP- and CTIP2-positive neuron (arrowhead) and tracing. Scale bar, 200 μm. **c**–**f**, No significant differences in dendritic segment numbers (**c**), number of branching points (**d**), dendritic spine density (**e**) and soma area (**f**) between TD, typical WS and pWS88 were observed. **g**–**m**, Morphometric analysis shown as individual participant for total dendritic length (**g**), dendritic tree number (**h**), dendritic spine number (**i**), dendritic segment number (**j**), number of branching points (**k**), dendritic spine density (**l**) and soma area (**m**). **n**, Four-week-old neurons were dissociated and plated to trace total neurite length every hour, for a total of 6 h. Representative images of traced neurons plated after 0 and 6 h from TD, typical WS and atypical pWS88 iPSC-derived neurons. **o**–**r**, Morphometric analysis showing significant differences among TD, typical WS and pWS88 in the initial neurite growth velocity (6 h period). **r**, Morphometric analysis shown for individual participants for neurite growth velocity for 6 h interval. *n*, Number of traced neurons. **s**–**u**, No significant changes were observed in the total dendritic length (**s**), dendritic segment number (**t**) and dendritic spine number (**u**) of TD neurons plated in different densities (300–1,200 cells per square millimetre). **v**, Individual channels of puncta quantification of post- and presynaptic markers (Homer1/Vglut1). Scale bar, 2 μm. For **c**–**m** and **o**–**u**, data are shown as mean ± s.e.m.; *n*, number of traced neurons, $^{*}P < 0.05$, $^{**}P < 0.01$, Kruskal–Wallis test (**c**–**f**), one-way ANOVA and Tukey's post hoc test (**o–q**, **r**–**u**).



i

Subject	Age	Sex	Postmortem interval (hours)	Time in formalin (years)	Cause of Death	Hemisphere
WS 1	31	M	26	22	Cardiorespiratory arrest	Left
WS 6	47	M	N/A	15	Died in sleep	Left
TD-C4	31	M	15	14	Moter vehicle accident	Left
TD-5250	51	M	25	5	Hypertensive cardiovascular disease	Left
TD-A02048	51	M	18	N/A	Myocardial infarction	Left
TD-C2	51	M	30	N/A	Myocardial infarction	Left

Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Alteration in calcium transient in WS iPSC-derived neurons and morphometric analysis of cortical layer V/VI pyramidal neurons in post-mortem tissue. a, Puncta quantification of post- and presynaptic markers. The synaptic proteins Vglut (presynaptic) and Homer1 (postsynaptic) were used as markers and only co-localized puncta on MAP2$^+$ cells were quantified and graphed. Data are shown as the mean ± s.e.m.; *n*, number of neurons. **b**, Summary of preparation of neurons for calcium transient analysis. Representative images of live neuronal culture expressing RFP driven by synapsin promoter and the uptake of Fluo-4AM calcium dye. **c**, Blockade of calcium transient by TTX inhibition of synaptic activity. **d**, Representative images of calcium transient in single neurons (RFP-positive, arrowhead) from TD (top), typical WS (middle) and pWS88 (bottom). Number in the lower right of each figure represents each time point (seconds) when change in Fluo-4AM occurs. **e**, **f**, Calcium transient analysis shown as individual for frequency (**e**) and percentage of signalling neurons (**f**). Data are shown as mean ± s.e.m.; *n*, number of fields analysed. **g**, MEA analyses revealed an increase in spontaneous neuronal spikes. Data show individual clones. **h**, Raster plot of TD and WS iPSC-derived neurons analysed by multi-electrode array. **i**, Details of individuals used for the analysis. **j**–**l**, No significant differences in dendrite number (**j**), dendritic spine density (**k**) and soma area (**l**) between TD and typical WS were observed. Data are shown as mean ± s.e.m.; *n*, number of traced neurons, two-sided unpaired Student's *t* test. **m**–**s**, Morphometric analysis shown for each individual for total dendritic length (**m**), dendritic spine number (**n**), segment number (**o**), branching point number (**p**), dendrite number (**q**), dendritic spine density (**r**) and soma area (**s**). Data are shown as mean ± s.e.m.; *n*, number of traced neurons.

Extended Data Table 1 | List of top ten most significant differentially expressed genes in WS compared with TD for NPC and neurons

NPC: TD x WS

Gene name	Description	Fold-change	p-value
SCN4A	sodium channel, voltage gated, type IV alpha subunit	-11.92	9.72E-10
SLC7A14	solute carrier family 7, member 14	-15.85	1.85E-12
SLC38A5	solute carrier family 38, member 5	-8.51	5.67E-09
ADGRA2	adhesion G protein-coupled receptor A2	-30.50	3.26E-08
SLC1A6	solute carrier family 1 (high affinity aspartate/glutamate transporter), member 6	8.03	1.56E-08
CXCL12	chemokine (C-X-C motif) ligand 12	-7.51	3.33E-08
SLC30A3	solute carrier family 30 (zinc transporter), member 3	10.28	8.32E-09
SLC8A2	solute carrier family 8 (sodium/calcium exchanger), member 2	-16.84	5.36E-13
HTR1B	5-hydroxytryptamine (serotonin) receptor 1B, G protein-coupled	-10.96	8.20E-09
SLC24A2	solute carrier family 24 (sodium/potassium/calcium exchanger), member 2	-13.25	4.43E-10

NPC: TD x pWS88

Gene name	Description	Fold-change	p-value
GABRA3	gamma-aminobutyric acid (GABA) A receptor, alpha 3	14.71	6.30E-10
SYT13	synaptotagmin XIII	6.55	1.87E-08
PPP2R2C	protein phosphatase 2, regulatory subunit B, gamma	15.74	7.69E-09
CELF4	CUGBP, Elav-like family member 4	7.27	1.22E-07
TRIM67	tripartite motif containing 67	15.08	6.52E-10
ADRA2A	adrenoceptor alpha 2A	8.44	1.42E-07
JAKMIP1	janus kinase and microtubule interacting protein 1	19.56	7.21E-09
CA10	carbonic anhydrase X	74.54	8.50E-10
LHFPL4	lipoma HMGIC fusion partner-like 4	8.31	5.71E-08
ACSL6	acyl-CoA synthetase long-chain family member 6	7.07	7.63E-08

Neuron: TD x pWS88

Gene name	Description	Fold-change	p-value
CRYM	crystallin, um	7.68	3.21E-05
RASL12	RAS-like, family 12	13.41	6.04E-08
PDLIM1	PDZ and LIM domain 1	5.44	1.00E-05
ZSCAN10	zinc finger and SCAN domain containing 10	58.80	6.22E-06
ANO1	anoctamin 1, calcium activated chloride channel	9.61	1.78E-06
DUSP23	dual specificity phosphatase 23	5.09	0.000114507
SLC16A5	solute carrier family 16 (monocarboxylate transporter), member 5	33.27	8.90E-05
KRT19	keratin 19, type I	9.94	0.000130985
TMEM30B	transmembrane protein 30B	17.09	2.11E-07
KIF18B	kinesin family member 18B	4.50	0.000217123

Neuron: TD x WS

Gene name	Description	Fold-change	p-value
FAM19A5	family with sequence similarity 19 (chemokine (C-C motif)-like), member A5	1076.39	2.23E-10
TPM2	tropomyosin 2 (beta)	-5.88	2.99E-08
SCN4A	sodium channel, voltage gated, type IV alpha subunit	-45.81	1.63E-07
IGSF21	immunoglobin superfamily, member 21	8.20	2.33E-07
TNNT2	troponin T type 2 (cardiac)	-29.87	3.58E-07
PXMP4	peroxisomal membrane protein 4, 24kDa	-7.39	3.83E-05
ZSCAN10	zinc finger and SCAN domain containing 10	35.64	9.74E-05
MYOZ1	filamin-, Actinin- And Telethonin-Binding Protein	-6.64	0.000119159
LAD1	ladinin 1	-24.81	0.000213205
PHOSPHO1	phosphatase, Orphan 1	33.95	0.000266061

Extended Data Table 2 | Most significant ($P < 0.05$) enriched GO terms in NPC of WS compared with TD samples

Down-regulated genes in WS compared to TD in NPC cells

GO TERM	GO Description	-Log (p-value)	P-value	Genes
GO:0005578	proteinaceous extracellular matrix	18.3452855	4.52E-19	WNT16,MMP25,DCN,LAMC3,COL16A1,WNT11,ADAMTS2,LAMB1,PAPLN,MMP9,WISP1,ECM2,COL1A1,CPZ,EFEMP1,LTBP2,COL10A1,COL21A1,OMD,ADAMTS8,WNT10A,EMILIN1,ADAMTS14,KERA,FBLN5,ADAMTS17,COL6A1,COL6A2,ADAMTS10,HMCN1,ADAMTS16,HAPLN1,ADAMTS4,COL6A3,COL1A2,MFAP4,SOST,NPNT,COL6A5,HPSE2,VWC2,PRELP,SPOCK3,LAMA2,ADAMTSL2,COL28A1
GO:0048856	anatomical structure development	16.80415844	1.57E-17	WNT16,MYLIP,TFAP2B,BAZ1B,CD4,DCN,SEMA3B,ADGRA2,RUNX3,IBSP,TG,ALX4,CAMK2B,TLE2,MAOB,RORA,EPHA8,TP63,PLXNA2,RARB,MEF2C,WNT11,TFAP2C,LAMB1,DSP,NEFH,MMP9,HCK,SGCG,FLT1,CRISPLD2,CPQ,MET,COBL,ENG,PITX3,COL1A1,CPT1A,CYP27B1,VDR,EYA4,TFEB,PCDHB2,PCDHB5,PCDHB6,PDGFRB,FN1,TFAP2E,LPPR4,RCAN3,TNNT2,BCL11A,PCDHB14,PCDHB12,INHBA,COL10A1,BMP2,TCF15,SIX1,HSPA2,PTPRB,CDH15,SERPINF1,LYVE1,PDGFRA,TBX3,HEY2,AGT,WNT10A,RAPGEF5,MYO7A,MSTN,ARHGAP24,CDH11,COL6A1,COL6A2,ITGA10,CASQ1,LEFTY2,EPHA5,HAPLN1,ARHGAP26,DCDC2,CACNA1C,EDNRA,UNC5D,ZIC3,MMP14,CACNA1D,NBL1,FGF17,ADAMTS4,STC1,ITGB2,PTH1R,PLXDC1,ALOX15,BRINP3,MSX1,FZD5,COL6A3,ELF3,SERPINI1,HEYL,PITX2,NPY1R,NPY5R,ITGA2,DACT2,COL1A2,SYK,DACT1,ANPEP,BATF2,NPNT,EFNA1,GPR183,BNC1,ALCAM,SIX2,SGCD,HAS2,PTGER4,NINJ2,SCG2,RARG,SSH3,ABLIM3,CSPG4,CMKLR1,LY6H,PCDHB9,TH,MAB21L1,GREM2,ADGRB1,FES,CSF1R,NTM,CAMK1D,GAS6,OPCML,SCN5A,TBX1,ALDH1A3,SLITRK6,SGCZ,TNFAIP2,NTF3,RTN4RL1,PDE2A,DNER,VWC2,PRELP,TDRD7,AKR1C3,RYR1,MME,LAMA2,PCDHB11,PAPSS2,MAFB,POU5F1,PCDHA7,PCDHA5,DPF3,ITGA1,GSTA1,PCDHA11,PCDHA10
GO:0007155	cell adhesion	16.70678997	1.96E-17	CD4,CLDN11,IBSP,TNC,LAMC3,RORA,EPHA8,COL16A1,LAMB1,FERMT1,HCK,SRPX,WISP1,ECM2,ENG,CXCL12,PRPH2,PCDHB2,PCDHB5,PCDHB6,FN1,PCDH17,PCDHB14,PCDHB12,ADGRE5,TNFAIP6,OMD,ADGRE2,ISLR,CDH15,MYBPH,LYVE1,PDGFRA,EMILIN1,PCDH10,FBLN5,CDH11,COL6A1,COL6A2,ITGA10,HAPLN1,CNTNAP5,ITGB2,AZGP1,SNED1,COL6A3,IGFBP7,ITGA2,EDIL3,SYK,MFAP4,NPNT,EFNA1,PCDH7,ALCAM,LRRN2,PTGER4,NINJ2,COL6A5,PCDHB9,ADGRB1,NTM,GAS6,OPCML,PKP3,THBS2,LAMA2,PCDHB11,ITGBL1,PCDHA7,PCDHA5,ITGA1,COL28A1,PCDHA13,PCDHGC3,ACTN3,PCDHA11,PCDHA10,PCDHGA7
GO:0030198	extracellular matrix organization	13.94978657	1.12E-14	DCN,IBSP,TNC,LAMC3,COL16A1,ADAMTS2,LAMB1,MMP9,ENG,COL1A1,EFEMP1,FN1,COL10A1,COL21A1,BMP2,EMILIN1,ADAMTS14,FBLN5,COL6A1,COL6A2,ITGA10,CTSK,HAPLN1,MMP14,ADAMTS4,ITGB2,CTSS,COL6A3,ITGA2,KLKB1,COL1A2,MFAP4,LTBP3,NPNT,HAS2,COL6A5,GAS6,LAMA2,ITGA1,COL28A1
GO:0005576	extracellular region	9.533946889	2.92E-10	WNT16,SCIN,MMP25,SYT7,DCN,SEMA3B,CLDN11,GPRC5A,ACPP,EHD2,IBSP,TNC,FAM65C,TG,LAMC3,ENTPD2,LY75,TLE2,BCL3,MAOB,FRMPD1,COL16A1,WNT11,ADAMTS2,LAMB1,DSP,SUSD2,APOL4,PAPLN,MMP9,BPI,MCF2,PLP2,ACP5,FLT1,WFDC1,CRISPLD2,CRYM,PDGFRL,CPQ,WISP1,NDRG1,MET,STX1A,PCOLCE,ECM2,ENG,PTGDS,CXCL12,PNPO,COL1A1,CPZ,OAS3,PDGFRB,POMC,EFEMP1,FN1,IL1R1,QPCT,ANGPTL1,VAMP8,LTBP2,ENOX1,CAT,CPXM2,XPNPEP2,INHBA,ADGRE5,COL10A1,EDN3,COL21A1,CPNE5,BMP2,PCSK2,CFP,HSPA2,OMD,TUBA4A,FGL2,ISLR,CDH15,DPP6,SERPINF1,GSTT2B,LYVE1,IAH1,ADAMTS8,DYSF,AGT,WNT10A,EMILIN1,ADAMTS14,MSTN,RBP5,KERA,SLC46A3,FBLN5,ADAMTS17,CDH11,COL6A1,COL6A2,NTN5,ADAMTS10,FCN3,HMCN1,CTSK,LEFTY2,TMEFF2,ADAMTS16,HAPLN1,GABRB2,PLA2G7,IGSF1,LCN9,SLC5A12,CCDC3,IGSF10,PLA2R1,TMPRSS11D,MMP21,GNA14,WIF1,NBL1,FGF17,ADAMTS4,STC1,C1R,ITGB2,PTH1R,AZGP1,FGFR4,PLXDC1,CXCL16,ITIH3,BRINP3,SNED1,HAAO,CTSS,S100A11,COL6A3,IGFBP7,SERPINI1,ALB,CDCP1,PRSS12,EDIL3,KLKB1,COL1A2,CHMP4C,AKR1E2,VWA2,MFAP4,ACSM1,ANPEP,GPT,SOST,LTBP3,NPNT,FSTL5,EFNA1,ALCAM,CYTL1,SOSTDC1,SCG2,COL6A5,HPSE2,CSPG4,ADGRG2,C1QTNF1,VPS37D,CLEC14A,EPS8L2,GREM2,PENK,C1S,GRID1,GAS6,OPCML,OLFML1,ALDH1A3,FAM19A3,MUC6,TNFAIP2,FAM212A,NTF3,RTN4RL1,THBS2,THSD4,VWC2,PRELP,PLAC9,SPOCK3,HLA-DRB1,AKR1C3,RYR1,MME,LAMA2,ADAMTSL2,SULT1C2,HLA-DRB5,ITGBL1,RASSF9,TGM2,ITGA1,COL28A1,APOL6,PCDHGC3,GSTA1,C4A,ACTN3,PCDHA10
GO:0005615	extracellular space	8.287607079	5.16E-09	WNT16,DCN,ACPP,IBSP,TNC,TG,TLE2,WNT11,LAMB1,APOL4,MMP9,FLT1,WFDC1,CPQ,WISP1,PCOLCE,ENG,PTGDS,COL1A1,OAS3,POMC,EFEMP1,FN1,ANGPTL1,LTBP2,ENOX1,CPXM2,INHBA,ADGRE5,EDN3,BMP2,PCSK2,CFP,HSPA2,FGL2,SERPINF1,AGT,WNT10A,MSTN,FBLN5,COL6A2,FCN3,CTSK,LEFTY2,PLA2G7,NBL1,FGF17,ADAMTS4,C1R,AZGP1,PLXDC1,CXCL16,CTSS,S100A11,COL6A3,IGFBP7,SERPINI1,ALB,KLKB1,COL1A2,VWA2,ACSM1,ANPEP,SOST,CYTL1,SOSTDC1,SCG2,C1QTNF1,C1S,GAS6,TNFAIP2,VWC2,SPOCK3,C4A
GO:0005886	plasma membrane	8.166494631	6.82E-09	CALCR,ABCC8,TBXA2R,SCN4A,CACNA2D2,MYLIP,MMP25,CD4,HFE,CLDN11,GPRC5A,ACPP,SLC38A5,ATP1A2,ADGRA2,EHD2,PTGER3,CYBA,ENTPD2,LY75,MCOLN3,CAMK2B,FRMPD1,CNGB1,EPHA8,MGLL,PLXNA2,PAG1,ATP8B1,IL12RB2,KCNK2,GABRP,DSP,KCNK6,SUSD2,SLC10A1,SLC8A3,FERMT1,HCK,BPI,PLP2,NALCN,SGCG,FLT1,NDRG1,MET,VIPR2,COBL,STX1A,ENG,RNF43,TCIRG1,TNS2,KCNA1,OAS3,ADGRD1,PCDHB2,PCDHB5,PCDHB6,SLC27A6,PDGFRB,IL1R1,LPPR4,SLC8A2,SGK1,VAMP8,PCDH17,GPR68,GRIA2,PCDHB14,PCDHB12,ENOX1,XPNPEP2,ADGRE5,PMEPA1,HSPA2,PTPRB,ADGRE2,ADCY4,CDH15,ATP1A4,LYVE1,TRPM1,PDGFRA,HTR1B,DYSF,SLC19A3,PCDH10,FBLN5,DISP2,CDH11,BEST4,ITGA10,RGS16,UNC80,SLC22A14,EPHA5,GABRB2,SLC22A3,HTR2C,SLC26A7,SLC5A12,PLCH2,LYPD6B,CACNA2D4,CACNA1C,ANO4,EDNRA,GABRA2,PLA2R1,TMPRSS11D,MCOLN2,FGD5,SLC24A2,GNA14,MMP14,CACNA1D,LYPD5,ABCG1,ITGB2,PTH1R,AZGP1,FGFR4,LY6E,PLXDC1,ALOX15,CXCL16,DHRS3,FZD5,CDCP1,MST1R,PRSS12,NPY1R,NPY5R,ITGA2,KLKB1,ACSL6,GRIK2,HCN1,SYK,ANPEP,PARM1,AFAP1L2,GSG1L,EFNA1,PTAFR,NPR1,GPR183,PCDH7,TM4SF1,SLC16A5,SIX2,SGCD,PRKCDBP,HAS2,KCND3,PTGER4,NINJ2,SYNPO,GPR22,HPSE2,EVC2,BNC2,CSPG4,ADGRG2,C1QTNF1,CMKLR1,SLCO3A1,LY6H,EPS8L2,PCDHB9,TH,GPR139,ADGRB1,FES,CSF1R,NTM,GRID1,RGS7,OPCML,SCN5A,PKP3,SLITRK6,TMEM173,ANKS1B,SGCZ,IFNLR1,PRKG1,EVI2B,RTN4RL1,PDE2A,TPCN1,QRFPR,NPSR1,GJA4,DNER,OR7D2,CLEC2A,NKAIN2,HLA-DRB1,RYR1,MME,CHRNG,PCDHB11,HLA-DRB5,PPAPDC1A,PCDHA7,PCDHA5,ITGA1,PCDHA13,PCDHGC3,C4A,FMN1,PCDHA11,PCDHA10,PCDHGA7
GO:0055085	transmembrane transport	7.872963268	1.34E-08	ABCC8,SCN4A,CACNA2D2,SLC7A14,SLC38A5,ATP1A2,MCOLN3,ATP2C2,CNGB1,ATP2A3,RARB,SLC25A43,ATP8B1,KCNK2,FXYD5,SLC17A6,GABRP,KCNK6,SLC10A1,SLC8A3,PLP2,NALCN,CPT1A,TCIRG1,KCNA1,SLC27A6,PEX5L,STEAP3,SLC8A2,SGK1,GRIA2,ADCY4,ATP1A4,TRPM1,ADAMTS8,SLC19A3,SLC46A3,BEST4,CASQ1,UNC80,SLC22A14,SLC25A48,GABRB2,SLC22A3,SLC26A7,SLC5A12,CACNA2D4,CACNA1C,ANO4,GABRA2,MCOLN2,SLC24A2,CACNA1D,ABCG1,AZGP1,COX7A1,ALB,GRIK2,HCN1,SLC16A5,KCND3,RARG,SLCO3A1,SLC17A8,SLC9A9,GRID1,SLC25A18,GAS6,SCN5A,PDE2A,TPCN1,RYR1,CHRNG

Up-regulated genes in WS compared to TD in NPC cells

GO TERM	GO Description	-Log (p-value)	P-value	Genes
GO:0007155	cell adhesion	2.964750421	0.00108455	ITGB5,TNR,TEK,PARVG,CDHR1,MPZL2,GRID2,ACAN,PIK3CD,EGFL7,CLDN6,GCNT1,CLDN4,PTPRT,DPP4
GO:0005886	plasma membrane	2.391582019	0.00405899	GABRA3,ERBB3,ITGB5,TNFRSF10A,SLC1A6,SLC6A12,GLP1R,SLC30A3,ABCG2,TEK,CSMD2,SDC4,BTN1A1,F12,PDE6B,MAP7,PARVG,RHCG,PPAP2C,ZNF185,PLIN2,CDHR1,KIRREL3,DLG2,GPR158,GRID2,ADCY8,XKR8,F11R,CLIC6,WNT4,OXGR1,OSCAR,PIK3CD,MRGPRF,GRM8,KCNB2,EPHA10,ADRA2C,CLDN6,CLDN4,PTPRT,DPP4,OCLN,CD247,HLA-DOA,SHISA9,HLA-DMB
GO:0006629	lipid metabolic process	2.063032094	0.00864904	ST3GAL6,ST8SIA5,SDC4,APOC1,CYP2J2,CYP11A1,PPAP2C,PLIN2,WNT4,HACD1,PIK3CD,ACOT12,FADS6,UGT8,SPTSSB
GO:0007009	plasma membrane organization	1.872296029	0.0134185	MAP7,XKR8,F11R,WNT4
GO:0006790	sulfur compound metabolic process	1.823703224	0.0150071	MGST1,ST3GAL6,SDC4,ACOT12,OPLAH,CHST6
GO:0007267	cell-cell signaling	1.725551179	0.0188126	GABRA3,PTPRN,SLC1A6,SLC6A12,TEK,DLG2,GRID2,ADCY8,KCNB2,MAFA,ADRA2C
GO:0048856	anatomical structure development	1.689882589	0.0204229	ERBB3,MYLK,ITGB5,HTATIP2,TNR,HPCAL4,PADI2,TEK,LIN28A,CHI3L1,TMOD1,RHCG,RP11-35N6.1,KIRREL3,MPZL2,DLG2,GRID2,ACAN,F11R,WNT4,HACD1,KRT19,PIK3CD,EGFL7,UGT8,ALX1,EPHA10,ADRA2C,CLDN4
GO:0005576	extracellular region	1.604486647	0.0248607	ST3GAL6,ERBB3,MYLK,ITGB5,ZDHHC15,CPVL,ENPP5,TNR,PADI2,TEK,NPPB,SDC4,BTN1A1,SPINK2,PRRG3,APOC1,F12,CHI3L1,CYP2J2,RHCG,CBLC,PLIN2,KIRREL3,ACAN,ITLN2,F11R,VWA5B1,CLIC6,WNT4,MUC3A,OSCAR,KRT19,EGFL7,MRGPRF,CREG2,EPHA10,BEX5,ADAMTSL5,PABPC1L2A,DPP4,GPX3,FAM19A5
GO:0007010	cytoskeleton organization	1.404009141	0.0394449	MYLK,ITGB5,CORO2A,MAP7,TMOD1,PARVG,F11R,CDC42BPG,KRT19
GO:0016757	transferase activity, transferring glycosyl groups	1.395223326	0.040251	ST3GAL6,ST8SIA5,GALNT14,UGT8,GCNT1

Extended Data Table 3 | Most significant ($P < 0.05$) enriched GO terms in neurons of WS compared with TD samples

Down-regulated genes in WS compared to TD in neurons

GO TERM	GO Description	-Log(p-value)	P-value	Genes
GO:0043167	ion binding	2.995842055	0.00100962	BAZ1B,FMO1,VRK2,DSG2,RFC2,ME1,ATP2A3,ADAMTS2,MYL9,TNNC2,ENO3,FOLR1,ACSS3,TRIM38,PCDHB5,TNNC1,ABCG2,CAT,ADGRE5,MYH2,ADGRE2,ZNF835,ATP8B3,MTL5,GSTM1,ACTA1,SH3RF2,ACTC1,ALOX15,CAPN13,ZNF283,ZNF558,ACOX2,GSTM4,SLFN12,RHOD,RNF212,ZNF626,SERPINA5,S100A13,S100A4,ZNF502,ZFP28,ZNF560,ZNF667,PEG3,ZNF726,ZNF737,ZNF578
GO:0008092	cytoskeletal protein binding	2.841323964	0.00144104	USH1C,MYBPC2,TNNC2,STX1A,TNNC1,TNNT2,MYH2,ACTA1,ACTC1,MYOZ1,TPM2
GO:0003013	circulatory system process	2.42626763	0.00374742	ELN,TNNC1,AGT,ACTC1,CD34
GO:0006950	response to stress	2.270514469	0.00536396	TBXA2R,BAZ1B,VRK2,RFC2,ATP2A3,IL4R,LAT2,MYL9,TRIM38,CAT,ADGRE5,MYH2,ADGRE2,MTL5,GSTM1,PARP9,ALOX15,ACOT11,BATF2,CD34,BHLHA15,SIGIRR,IFITM2,SERPINA5,HLA-DRB1,HLA-DRB5,POU5F1
GO:0005198	structural molecule activity	2.015682009	0.00964535	MYOM2,ELN,MYBPC2,MYL9,MYH2,TINAGL1,ACTA1,LAD1,TPM2
GO:0005829	cytosol	1.675077895	0.0211311	USH1C,ME1,MYBPC2,MYL9,TNNC2,STX1A,EIF4H,ENO3,TRIM38,TNNC1,TNNT2,MYH2,GSTM1,PARP9,ACTA1,ACTC1,ALOX15,GSTM4,RHOD,ALDH1A3,S100A13,TPM2,POU5F1
GO:0005777	peroxisome	1.574933909	0.0266113	PXMP4,CAT,ACOX2
GO:0003674	Molecular function	1.54856974	0.0282768	USH1C,TBXA2R,SCN4A,BAZ1B,FMO1,ACPP,VRK2,MYOM2,DSG2,PREX2,ELN,RFC2,ME1,ATP2A3,IL4R,LAT2,MYBPC2,ADAMTS2,PROKR2,MYL9,PXMP4,TNNC2,STX1A,TBL2,EIF4H,ENO3,FOLR1,ACSS3,TRIM38,SLC22A2,PCDHB5,SLC27A6,TNNC1,TNNT2,ABCG2,CAT,ADGRE5,MYH2,ADGRE2,ZNF835,ATP8B3,NSUN5,MTL5,DMGDH,GSTM1,AGT,THUMPD2,PARP9,TINAGL1,ACTA1,RASSF3,SH3RF2,LAD1,ACTC1,ALOX15,ACOT11,UBXN10,CAPN13,RBM47,TC2N,ZNF283,ZNF558,BATF2,ACOX2,GSTM4,CYTL1,PRKCDBP,SLFN12,RHOD,CD34,CHRNA9,GCNT4,MYOZ1,RNF212,BHLHA15,MAATS1,ALDH1A3,RNLS,RBM43,SIGIRR,CCK,ZNF626,LRRIQ4,SERPINA5,S100A13,HLA-DRB1,S100A4,ZNF502,ZFP28,SLC2A10,ZNF560,ZNF667,PEG3,TPM2,HLA-DRB5,LAYN,POU5F1,ZNF726,ANKRD65,ZNF737,ZNF578,SLC22A31,TCF24
GO:0016765	transferase activity, transferring alkyl or aryl (other than methyl) groups	1.475366347	0.0334683	GSTM1,GSTM4
GO:0016887	ATPase activity	1.459248951	0.0347337	RFC2,ATP2A3,TNNT2,ABCG2,ATP8B3,ACTC1
GO:0016874	ligase activity	1.406285389	0.0392387	RFC2,ACSS3,TRIM38,SLC27A6,SH3RF2,RNF212
GO:0015979	photosynthesis	1.387345392	0.0409878	RFC2
GO:0005615	extracellular space	1.320715681	0.0477842	ACPP,ENO3,ADGRE5,AGT,TINAGL1,ACTA1,ACTC1,CYTL1,SERPINA5,S100A13,S100A4

Up-regulated genes in WS compared to TD in neurons

GO TERM	GO Description	-Log(p-value)	P-value	Genes
GO:0001071	nucleic acid binding transcription factor activity	2.409059017	0.00389889	PITX1,TFAP2C,HAND1,NR5A2,IRF6,ZSCAN10,KLF4,ALX1,DMBX1,FOXB2
GO:0008233	peptidase activity	1.531733923	0.0293945	HPN,PRSS16,RHBDF2,ADAMTS16,USP41,TMPRSS2
GO:0009790	embryo development	1.354676981	0.0441899	HPN,HAND1,NR5A2,KLF4
GO:0006091	generation of precursor metabolites and energy	1.330439536	0.0467262	ALDOC,MT-ND2,MT-ND4,MT-ND1
GO:0048856	anatomical structure development	1.312454095	0.0487019	PITX1,TFAP2C,HPN,ALDOC,HAND1,NR5A2,IRF6,ZSCAN10,KLF4,PRDM14,BMP6,ISL2,PROK2,PHOSPHO1,ALX1,DMBX1

doi:10.1038/nature18958

Asymmetric division of contractile domains couples cell positioning and fate specification

Jean-Léon Maître[1]†, Hervé Turlier[1]*, Rukshala Illukkumbura[1]*, Björn Eismann[1]†, Ritsuya Niwayama[1], François Nédélec[1] & Takashi Hiiragi[1]

During pre-implantation development, the mammalian embryo self-organizes into the blastocyst, which consists of an epithelial layer encapsulating the inner-cell mass (ICM) giving rise to all embryonic tissues[1]. In mice, oriented cell division, apicobasal polarity and actomyosin contractility are thought to contribute to the formation of the ICM[2–5]. However, how these processes work together remains unclear. Here we show that asymmetric segregation of the apical domain generates blastomeres with different contractilities, which triggers their sorting into inner and outer positions. Three-dimensional physical modelling of embryo morphogenesis reveals that cells internalize only when differences in surface contractility exceed a predictable threshold. We validate this prediction using biophysical measurements, and successfully redirect cell sorting within the developing blastocyst using maternal myosin (*Myh9*)-knockout chimaeric embryos. Finally, we find that loss of contractility causes blastomeres to show ICM-like markers, regardless of their position. In particular, contractility controls Yap subcellular localization[6], raising the possibility that mechanosensing occurs during blastocyst lineage specification. We conclude that contractility couples the positioning and fate specification of blastomeres. We propose that this ensures the robust self-organization of blastomeres into the blastocyst, which confers remarkable regulative capacities to mammalian embryos.

During the 8- to 16-cell stage transition, oriented divisions can push one of the daughter cells towards the inside of the embryo[5,7,8]. Alternatively, blastomeres were observed to internalize after the 8- to 16-cell stage division, possibly driven by differences in cell contractility[2,3]. How the embryo generates cell populations with distinct contractile properties is, however, unknown, and the physical mechanism by which this leads to internalization remains disputed[2,3].

We first investigated the origin of the differences in contractility among blastomeres at the 16-cell stage. During the 8-cell stage, blastomeres polarize by forming an apical domain, which occupies only a portion of the contact-free surface. This apical material can be asymmetrically inherited during the following division[2,9], giving rise to both polarized and unpolarized blastomeres within the 16-cell-stage embryo, as can be observed from the levels of the essential apical protein aPKC[4,10,11] (Extended Data Fig. 1). We observe that unpolarized blastomeres, with low levels of aPKC, show higher cortical levels of myosin than polarized ones (Extended Data Fig. 1). Moreover, in embryos knocked out for two isoforms of aPKC[4,10], we observe no reduction of myosin levels where the apical domain would normally be, suggesting that aPKC antagonizes cortical myosin accumulations at the apical domain (Extended Data Fig. 1). As a consequence of reduced levels of myosin, we would expect the apical domain to exhibit reduced contractility. To test this, we took advantage of the periodic contractions that appear during the 8-cell stage[12,13] and used them as a proxy of contractility. Upon polarization of 8-cell-stage blastomeres, we measured contractions of lower amplitude at the apical domain than in the rest of the cortex (apical/non-apical: 59 ± 23%, mean ± standard deviation (s.d.), $n = 17$ blastomeres; Fig. 1a–e and Supplementary Video 1). After asymmetric division of 8-cell-stage blastomeres, the polarized 16-cell-stage blastomeres often show no detectable periodicity (52% of 23 polarized blastomeres showing contractions; Fig. 1d and Supplementary Video 2) and their contractions display lower amplitudes than those of unpolarized blastomeres (polarized/unpolarized: 65 ± 26%, mean ± s.d., $n = 23$ doublets; Fig. 1f–i). More precisely, we find that this difference in contractility between polarized and unpolarized blastomeres intensifies as cells internalize (Extended Data Fig. 2). Contractility in polarized blastomeres remains dampened by the apical domain (Fig. 1j and Extended Data Figs 1–2) and polarized sister cells resulting from symmetric divisions show little cortical heterogeneity and do not internalize (Extended Data Fig. 2). By contrast, unpolarized blastomeres increase their contractility, as initiated during compaction at the 8-cell stage[12], resulting in increasing heterogeneity in doublets stemming from asymmetric divisions (Fig. 1j and Extended Data Fig. 2). In summary, we identify the asymmetric inheritance of the apical domain during the 8- to 16-cell-stage division as the source of differences in contractility among blastomeres.

How these differences in cell contractility physically translate into the internalization of the ICM remains unclear[2,3]. To describe quantitatively the mechanism of internalization, we considered the blastomere surface tensions, which are controlled by actomyosin contractility[14], and used them to build a physical model of blastomere configuration. First, we considered a cell doublet as a minimal system in which one cell envelops its neighbour in an entosis-like process[15] (Extended Data Fig. 2 and Supplementary Video 5). This reductionist approach is justified by the fact that doublets resulting from asymmetrically divided 8-cell-stage blastomeres recapitulate both the morphogenesis and fate specification of the whole embryo[16,17]. Noting γ_c, the surface tension at cell–cell contacts, and γ_i, the tension at the cell-medium interface of the cell i (with $i = 1$ or 2 for a cell doublet; Fig. 2a), we define three dimensionless parameters: a compaction parameter $\alpha = \gamma_c/2\gamma_2$; a volume asymmetry $\beta = (V_1/V_2)^{1/3}$, where V_1 and V_2 are the volumes of each cell; and a tension asymmetry $\delta = \gamma_1/\gamma_2$. We can analytically derive the conditions for cell internalization (Fig. 2b, c, Supplementary Video 3 and Supplementary Note). Full internalization occurs whenever $\delta > 1 + 2\alpha$ (Fig. 2b, c), thus defining an internalization threshold $\delta_c = 1 + 2\alpha$ for the tension asymmetry, in agreement with previous numerical studies[18–20]. Before this transition, partial internalization configurations are predicted, which match the configurations observed experimentally in doublets of 16-cell-stage blastomeres (Extended Data Fig. 2 and Supplementary Video 5). Interestingly, the internalization threshold δ_c is not influenced by the size asymmetry β but depends critically on the compaction parameter α (Extended Data Fig. 3). Modulating α in the absence of tension asymmetry is, however, not sufficient for driving

[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany. †Present addresses: Mechanics of Mammalian Development Group, Institut Curie, CNRS UMR 3215, INSERM U934, 26, rue d'Ulm, 75248 Paris Cedex 05, France (J.-L.M.); Bioquant, Im Neuenheimer Feld 267, 69120 Heidelberg, Germany (B.E.).
*These authors contributed equally to this work.

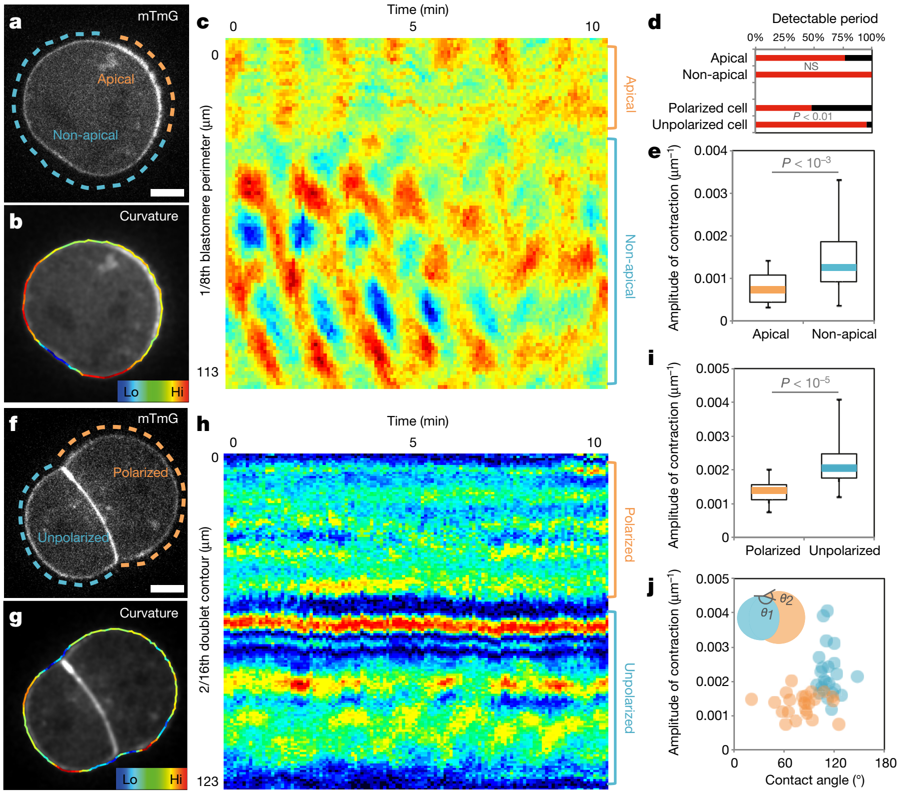


Figure 1 | Asymmetric inheritance of the apical domain generates blastomeres of different contractility. a–c, Eight-cell-stage blastomere expressing mTmG (**a**) with colour-coded surface curvature (**b**) and corresponding kymograph (**c**). Apical domain is highlighted in orange and non-apical cortex in blue. **d,** Proportion of blastomeres for which a contraction period can be detected (17 blastomeres and 23 doublets from 4 and 5 experiments, respectively). Mann–Whitney *U*-test *P* value; NS, not significant. **e,** Box plot of contraction amplitudes for apical (orange) and non-apical cortex (blue). Seventeen blastomeres from four experiments, Student's *t*-test *P* value. **f–h,** Doublet of 16-cell-stage blastomeres expressing mTmG (**f**) with colour-coded surface curvature (**g**) and corresponding kymograph (**h**). Polarized blastomere is highlighted in orange, unpolarized one in blue. **i,** Box plot of contraction amplitudes for polarized (orange) and unpolarized blastomeres (blue). Twenty-three doublets from four experiments. Student's *t*-test *P* value. **j,** Amplitude of contractions as a function of the contact angles θ_1 for polarized (orange) and θ_2 for unpolarized blastomeres (blue, Pearson $R = -0.611$, $n = 46$ blastomeres from five experiments, $P < 0.001$). Scale bars, 10 μm.

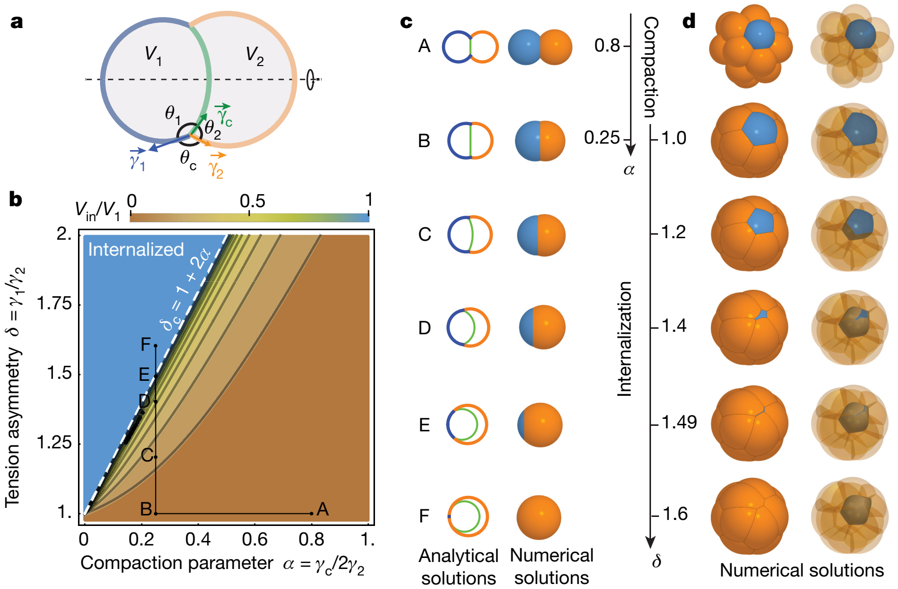


Figure 2 | Physical model of cell internalization. a, Schematic diagram of a cell doublet with surface tensions γ_1, γ_2 and γ_c of cell 1 (blue), 2 (orange) and the contact (green), respectively. Contact angles θ_1, θ_2 and θ_c and cell volumes V_1 and V_2 are also shown. **b,** Phase diagram describing the mechanical equilibrium of a doublet as a function of the compaction parameter α and tension asymmetry δ. Colour-coded degree of internalization (measured as the relative volume of cell 1 that is internalized V_{in}/V_1), threshold value δ_c at which internalization occurs (white dotted line) and an example of compaction (A to B) followed by internalization (B to F) in black (Supplementary Video 3–4) are overlaid. **c–d,** Analytical (left) and numerical solutions (right) for a doublet (**c**). Numerical solutions (**d**) for the compaction of 16 cells with opaque (left) and transparent (right) non-internalizing cells. The compaction parameter α decreases from 0.8 to 0.25, followed by an increase of tension asymmetry δ from 1.0 to 1.6.

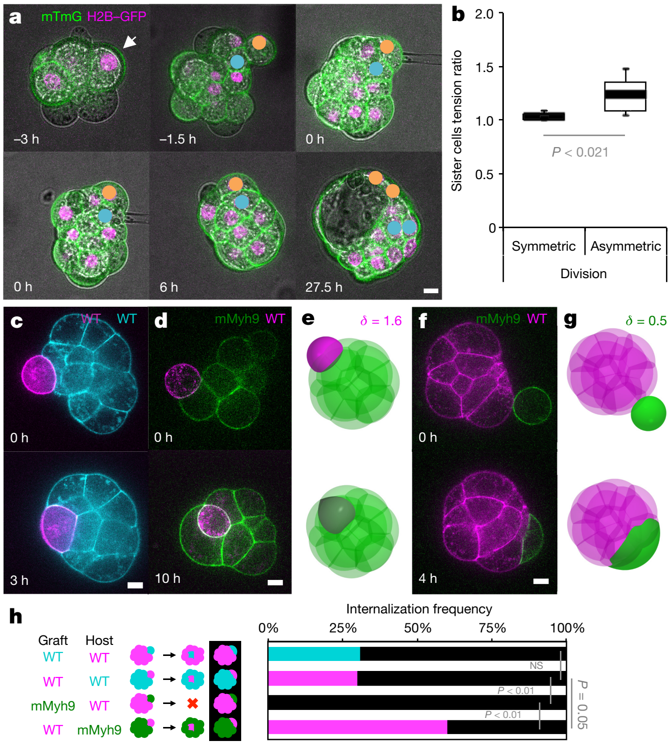


Figure 3 | Tension heterogeneities drive cell sorting of the ICM. **a**, Lineage tracking of polarized (yellow) and unpolarized (blue) daughter cells after surface tension measurement of mTmG (green) and H2B–GFP (magenta) expressing embryos. **b**, Box plot of surface tension ratio for sister cells with (symmetric) or without (asymmetric) internalization of one of the sister cell. Eight and seven pairs of cells from eleven embryos from five experiments, Student's *t*-test *P* values. **c**–**g**, Wild-type (WT; magenta or cyan) or mMyh9 (green) blastomeres grafted onto host embryos (**c**, **d**, **f**). Simulations of grafting experiments: one cell with $\delta = 1.6$ (**e**), corresponding to wild type onto mMyh9 (**d**); one cell with $\delta = 0.5$ (**g**), corresponding to mMyh9 onto wild type (**f**). **h**, Internalization frequencies for chimaeric embryos (wild type–wild type (13 mG host embryos and 20 mTmG host embryos from 3 experiments) and wild type–mMyh9 (12 mG host embryos and 20 mMyh9 host embryos from 4 experiments)). Black indicates that no internalization occurs, Mann–Whitney *U*-test *P* values; NS, not significant. Scale bars, 10 μm.

internalization. For the value of the compaction parameter measured at late 8-cell stage[12], $\alpha \approx 0.25$, we predict that any tension asymmetry δ higher than $\delta_c \approx 1.5$ should lead to complete internalization (Fig. 2b, c). Therefore, when measuring tension asymmetries, we expect that δ should not exceed ~1.5, otherwise the cell should be fully internalized and hence inaccessible to non-invasive methods.

To generalize this approach to the formation of the ICM in an embryo with 16 cells, we built a three-dimensional numerical model of the embryo using a multi-material mesh-based surface-tracking method[21] (Supplementary Note). We find in simulated embryos the same internalization conditions as in cell doublets, with a transition occurring above the threshold value $\delta_c \approx 1.5$ (Fig. 2d, Extended Data Fig. 7 and Supplementary Video 4). Therefore, the same physical mechanisms can explain the envelopment of one cell by another[15] and the sorting of cells within a tissue[18,19,22,23].

To test the predictions of the model experimentally, we used microaspiration to measure the surface tensions of sister cells after the 8- to 16-cell-stage division, visualize their internalization and then track their position within the blastocyst (Fig. 3a and Supplementary Video 6). Sister cells remaining at the surface of the embryo and failing to contribute to the ICM in the blastocyst show no tension asymmetry at the 16-cell stage ($\delta = 1.04 \pm 0.03$, mean ± s.d., $n = 8$ pairs of sister cells; Fig. 3b). On the other hand, when one sister cell internalizes after the 8- to 16-cell-stage division and contributes to the ICM within the blastocyst, we measure a tension asymmetry of 1.24 ± 0.17 (mean ± s.d., $n = 7$ pairs of sister cells; Fig. 3b). This confirms that asymmetric divisions are the source of tension heterogeneities in the embryo. Furthermore, the observation that internalizing cells are the only ones showing tension heterogeneity relative to their sister cells supports the hypothesis that tension heterogeneity is sufficient to drive cell internalization. Finally, the measured tension asymmetries are indeed lower than the internalization threshold value $\delta_c \approx 1.5$, as predicted by our theory.

To test directly the proposed internalization mechanism, we first generated embryos that lack the maternal allele of *Myh9* (ref. 24; mMyh9 hereafter), the specific isoform of myosin heavy chain that is required for pre-implantation development (zygotic[25] and maternal zygotic Myh10-knockout embryos form normal blastocysts; Supplementary Video 7). Although compaction is delayed due to their inability to generate sufficient tensions[12] (Extended Data Fig. 4), these embryos form blastocysts (Supplementary Video 8) and viable offspring. Next, we transplanted onto a mMyh9 host embryo a wild-type blastomere, which typically internalizes (60% of 20 grafted blastomeres internalized; Fig. 3d, h and Supplementary Video 9; for corresponding numerical simulation, see Fig. 3e and Supplementary Video 10) and contributes to the ICM of the host embryo. By contrast, when transplanted onto wild-type embryos, an mMyh9 blastomere always remains at the surface of the embryo (none of 12 grafted blastomeres internalized; Fig. 3e, h and Supplementary Video 11; for corresponding numerical simulation, see Fig. 3g and Supplementary Video 12), and stretches to envelop blastomeres of the host embryo. In comparison, transplanting wild-type cells onto wild-type hosts leads to a lower internalization frequency than with mMyh9 hosts (31% of 33 blastomeres internalized, $P = 0.05$; Fig. 3c, h and Supplementary Video 13). We conclude that the hierarchy of contractility between blastomeres is sufficient to direct cell internalization. As internalizing cells do not have an apical domain (Fig. 1 and Extended Data Figs 1–2) and do not require intact contractility of neighbouring cells (Fig. 3d), the mechanism by which cells internalize is analogous to a cell sorting process[22] and is distinct from an apical constriction, as previously proposed[3].

While cells adopt their position within the embryo, they segregate into two distinct lineages: trophectoderm and ICM. In the mouse embryo, this lineage specification is regulated by Yap subcellular localization[4,11], which, in cultured cells, is regulated by contractile forces[6]. Therefore, we tested whether contractility influences Yap localization and thereby fate specification. In agreement with previous studies[2,4], we find less cytoplasmic phosphorylated Yap in cells closer to the surface than in the internalized ones (Extended Data Fig. 5a, h). These blastomeres initiate trophectoderm specification as their surface cells show the highest Cdx2 levels (Extended Data Fig. 5a, o). In embryos treated with different concentrations of the myosin inhibitor blebbistatin (Bb), the correlation between these fate markers and cell position is weakened in a dose-dependent manner (Extended Data Figs 5–6). Moreover, mMyh9 embryos show similar defects at the 16-cell stage (Extended Data Figs 5g, n, u and 6c, f), despite being viable (unlike Bb-treated embryos). While inhibiting contractility de-compacts embryos[12], the position of cells is not shuffled, and yet the localization of Yap is affected. During cell internalization, outer cells deform extensively. This is especially the case for doublets in which, similarly to complete embryos, cytoplasmic localization of phosphorylated Yap is lowest for outer cells (Fig. 4a, c, d). When cell deformation is blocked by Bb treatment, phosphorylated Yap localization and Cdx2 levels become homogeneous within doublets (Fig. 4b–f). Remarkably, inhibition of contractility causes both blastomeres to become inner-cell-like with respect to phosphorylated Yap localization and Cdx2 levels, despite their external position (Fig. 4). This is consistent with internalizing cells reducing their contractility over their entire surface, since cell–cell contacts have a contractility equivalent to Bb-treated cells[12]. Together, these results indicate that without contractile forces, blastomeres adopt an inner-cell-like fate,

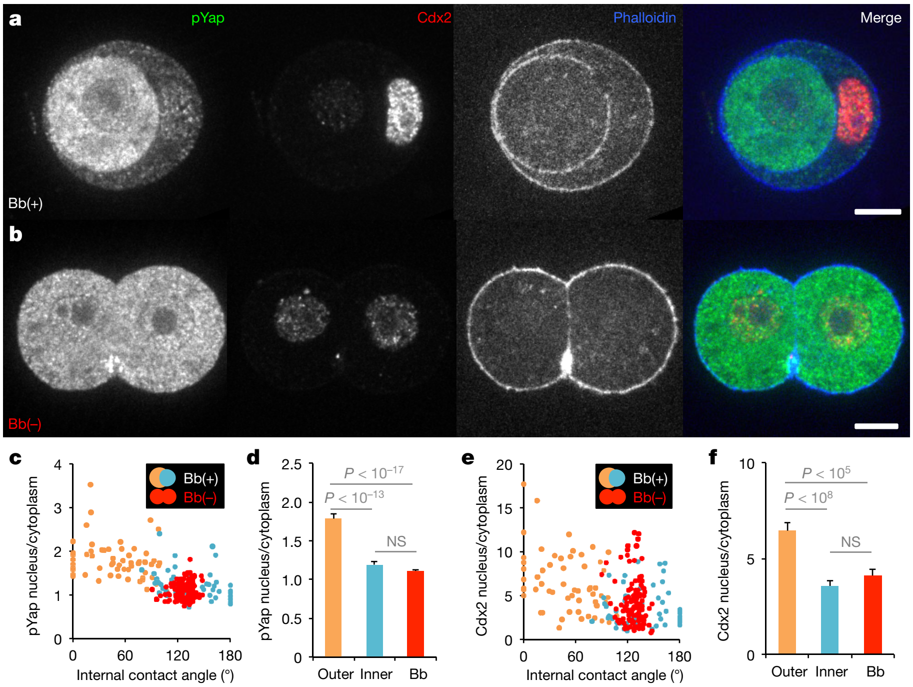


Figure 4 | Contractility couples morphogenesis and fate specification. **a**, **b**, Immunostaining of doublets of 16-cell-stage blastomeres treated with 25 μM Bb(+) (**a**; an inactive enantiomere of the inhibitor) or Bb(−) (**b**; the selective inhibitor of myosin II ATPase activity) showing phosphorylated (p)Yap (green), Cdx2 (red) and phalloidin (blue). **c**–**f**, Nucleus-to-cytoplasm intensity ratio of pYap (**c**) or Cdx2 (**e**) as a function of the internal contact angle for doublets treated with Bb(+) (outer cells in orange and inner cells in blue, pYap $R = -0.630$, or Cdx2 $R = -0.493$, $n = 59$ doublets from 3 experiments, $P < 0.001$) or Bb(−) (red, pYap $R = 0.158$, or Cdx2 $R = -0.118$, $n = 60$ doublets from 3 experiments, $P > 0.1$). Mean ± standard error of the mean (s.e.m.) nucleus to cytoplasm intensity ratio of pYap (**d**) or Cdx2 (**f**). For doublets treated with Bb(+), outer cells are shown in orange and inner cells in blue, while Bb(−)-treated cells are in red. Student's *t*-test *P* values; NS, not significant.

regardless of their position. Therefore, disrupting contractility uncouples morphogenesis and fate specification. Moreover, the control of Yap subcellular localization by contractile forces, reminiscent of those from cell culture studies[6], raises the possibility that lineage specification in the blastocyst could be mechanosensitive. Whether cells may sense their macroscopic deformation[26] or stresses at the molecular level[27], will require further studies. Such mechanosensitivity could explain why inner-cell-like localization of Yap is often observed for blastomeres before they have completed their internalization[2,4] (Extended Data Figs 5–6). We propose that the coupling of cell positioning and fate specification by contractility enables blastomeres to anticipate their final position and initiate their differentiation accordingly.

Apicobasal polarity can control the position of blastomeres by orienting the 8- to 16-cell-stage divisions relative to the surface of the embryo[5]. However, oriented cell divisions do not guarantee that internalized cells will be maintained inside the embryo. We find that apicobasal polarity also controls internalization by maintaining low contractility at the apical domain (Fig. 1). This allows unpolarized blastomeres to outcompete their polarized neighbours when their contractility grows above a threshold value (Figs 2 and 3). The resulting cell sorting is a fail-safe mechanism to one-shot oriented cell divisions. Such complementary mechanisms, together with the ability of cells to read mechanical cues[6] to guide their differentiation[28] (Fig. 4), can account for the regulative capacity of early mammalian embryos[7]. From this study, a self-organization framework determining the initial steps of morphogenesis and lineage specification in mammalian embryos is emerging.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 12 February; accepted 23 June 2016.
Published online 3 August 2016.

1. Wennekamp, S., Mesecke, S., Nédélec, F. & Hiiragi, T. A self-organization framework for symmetry breaking in the mammalian embryo. *Nature Rev. Mol. Cell Biol.* **14,** 452–459 (2013).
2. Anani, S., Bhat, S., Honma-Yamanaka, N., Krawchuk, D. & Yamanaka, Y. Initiation of Hippo signaling is linked to polarity rather than to cell position in the pre-implantation mouse embryo. *Development* **141,** 2813–2824 (2014).
3. Samarage, C. R. *et al.* Cortical tension allocates the first inner cells of the mammalian embryo. *Dev. Cell* **34,** 435–447 (2015).
4. Hirate, Y. *et al.* Polarity-dependent distribution of angiomotin localizes Hippo signaling in preimplantation embryos. *Curr. Biol.* **23,** 1181–1194 (2013).
5. Dard, N., Louvet-Vallée, S. & Maro, B. Orientation of mitotic spindles during the 8- to 16-cell stage transition in mouse embryos. *PLoS ONE* **4,** e8171 (2009).
6. Dupont, S. *et al.* Role of YAP/TAZ in mechanotransduction. *Nature* **474,** 179–183 (2011).
7. Johnson, M. H. From mouse egg to mouse embryo: polarities, axes, and tissues. *Annu. Rev. Cell Dev. Biol.* **25,** 483–512 (2009).
8. Watanabe, T., Biggins, J. S., Tannan, N. B. & Srinivas, S. Limited predictive value of blastomere angle of division in trophectoderm and inner cell mass specification. *Development* **141,** 2279–2288 (2014).
9. Johnson, M. H. & Ziomek, C. A. The foundation of two distinct cell lineages within the mouse morula. *Cell* **24,** 71–80 (1981).
10. Matsumoto, M. *et al.* PKCλ in liver mediates insulin-induced SREBP-1c expression and determines both hepatic lipid content and overall insulin sensitivity. *J. Clin. Invest.* **112,** 935–944 (2003).
11. Hirate, Y. *et al.* Par-aPKC-dependent and -independent mechanisms cooperatively control cell polarity, Hippo signaling, and cell positioning in 16-cell stage mouse embryos. *Dev. Growth Differ.* **57,** 544–556 (2015).
12. Maître, J.-L., Niwayama, R., Turlier, H., Nédélec, F. & Hiiragi, T. Pulsatile cell-autonomous contractility drives compaction in the mouse embryo. *Nature Cell Biol.* **17,** 849–855 (2015).
13. Lehtonen, E. Changes in cell dimensions and intercellular contacts during cleavage-stage cell cycles in mouse embryonic cells. *J. Embryol. Exp. Morphol.* **58,** 231–249 (1980).
14. Heisenberg, C.-P. & Bellaïche, Y. Forces in tissue morphogenesis and patterning. *Cell* **153,** 948–962 (2013).
15. Overholtzer, M. *et al.* A nonapoptotic cell death process, entosis, that occurs by cell-in-cell invasion. *Cell* **131,** 966–979 (2007).
16. Dietrich, J.-E. & Hiiragi, T. Stochastic patterning in the mouse pre-implantation embryo. *Development* **134,** 4219–4231 (2007).
17. Johnson, M. H. & Ziomek, C. A. Cell interactions influence the fate of mouse blastomeres undergoing the transition from the 16- to the 32-cell stage. *Dev. Biol.* **95,** 211–218 (1983).

18. Graner, F. & Glazier, J. A. Simulation of biological cell sorting using a two-dimensional extended Potts model. *Phys. Rev. Lett.* **69,** 2013–2016 (1992).
19. Brodland, G. W. The Differential Interfacial Tension Hypothesis (DITH): a comprehensive theory for the self-rearrangement of embryonic cells and tissues. *J. Biomech. Eng.* **124,** 188–197 (2002).
20. Guzowski, J., Korczyk, P. M., Jakiela, S. & Garstecki, P. The structure and stability of multiple micro-droplets. *Soft Matter* **8,** 7269–7278 (2012).
21. Da, F., Batty, C. & Grinspun, E. Multimaterial mesh-based surface tracking. *ACM Trans. Graph.* **33,** 112 (2014).
22. Krieg, M. *et al.* Tensile forces govern germ-layer organization in zebrafish. *Nature Cell Biol.* **10,** 429–436 (2008).
23. Maître, J.-L. *et al.* Adhesion functions in cell sorting by mechanically coupling the cortices of adhering cells. *Science* **338,** 253–256 (2012).
24. Jacobelli, J. *et al.* Confinement-optimized three-dimensional T cell amoeboid motility is modulated via myosin IIA-regulated adhesions. *Nature Immunol.* **11,** 953–961 (2010).
25. Wang, A. *et al.* Nonmuscle myosin II isoform and domain specificity during early mouse development. *Proc. Natl Acad. Sci. USA* **107,** 14645–14650 (2010).
26. Aragona, M. *et al.* A mechanical checkpoint controls multicellular growth through YAP/TAZ regulation by actin-processing factors. *Cell* **154,** 1047–1059 (2013).
27. Benham-Pyle, B. W., Pruitt, B. L. & Nelson, W. J. Mechanical strain induces E-cadherin-dependent Yap1 and β-catenin activation to drive cell cycle entry. *Science* **348,** 1024–1027 (2015).
28. Shin, J.-W. *et al.* Contractile forces sustain and polarize hematopoiesis from stem and progenitor cells. *Cell Stem Cell* **14,** 81–93 (2014).

Supplementary Information is available in the online version of the paper.

Acknowledgements We are grateful to the Hiiragi laboratory members and the European Molecular Biology Laboratory (EMBL) animal facility for their support. We thank F. Da and C. Batty for discussions on the simulations. We thank Y. Bellaïche for comments on an earlier version of the manuscript. Marie Curie individual fellowships under FP7 and H2020 programs support J.-L.M., H.T. and R.N. under Research Executive Agency grant agreements 329044, 656306 and 326701, respectively. H.T. acknowledges support from the Bettencourt-Schueller and Joachim Herz foundations. The Hiiragi laboratory is supported by EMBL, the European Research Council and VolkswagenStifftung.

Author Contributions J.-L.M. designed the project and experiments, and wrote the manuscript with input from all authors. J.-L.M. and B.E. performed and analysed the tension and lineage mapping experiments. J.-L.M. and R.I. performed and analysed the remaining experiments. R.N. helped with image analysis of the periodic contractions. H.T. designed the physical model and performed the simulations with help from F.N. T.H. supervised the study and helped design the project.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to J.-L.M. (jean-leon.maitre@curie.fr) or T.H. (hiiragi@embl.de).

Reviewer Information *Nature* thanks D. Discher, P.-F. Lenne, B. Plusa and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Embryo work. *Recovery and culture.* All animal work was performed in the animal facility at the European Molecular Biology Laboratory, with permission from the institutional veterinarian overseeing the operation (ARC number TH11 00 11). The animal facilities are operated according to international animal welfare rules (Federation for Laboratory Animal Science Associations guidelines and recommendations).

Embryos are isolated from superovulated female mice mated with male mice. Superovulation of female mice is induced by intraperitoneal injection of 5 international units (IU) of pregnant mare's serum gonadotropin (PMSG; Intervet Intergonan), followed by intraperitoneal injection of 5 IU human chorionic gonadotropin (hCG; Intervet Ovogest 1500) 44–48 h later. Two-cell-stage (embryonic day 1.5 (E1.5)) embryos are recovered by flushing oviducts from plugged females with 37 °C FHM (Millipore, MR-024-D) using a custom-made syringe (Acufirm, 1400 LL 23).

Embryos are handled using an aspirator tube (Sigma, A5177-5EA) equipped with a glass pipette pulled from glass micropipettes (Blaubrand intraMark).

Embryos are placed in KSOM (Millipore, MR-121-D) or FHM supplemented with 0.1% BSA (Sigma, A3311) in 10 μl droplets covered in mineral oil (Sigma, M8410 or Acros Organics). Embryos are cultured in an incubator with a humidified atmosphere supplemented with 5% CO_2 at 37 °C.

For imaging, embryos are placed in 5 cm glass-bottom dishes (MatTek).

Mouse lines. (C57BL/6xC3H) F1 hybrid strain is used for wild type (WT).

To visualize filamentous actin, LifeAct–GFP mice (*Tg(CAG-EGFP)#Rows*) are used[29]. To visualize plasma membranes, mTmG mice (*Gt(ROSA)26Sor*$^{tm4(ACTB\text{-}tdTomato,\text{-}EGFP)Luo}$) are used[30]. To visualize nuclei, H2B–GFP mice are used[31]. Genes are deleted maternally using Zp3-cre (*Tg(Zp3-cre)93Knw*) mice[32]. To generate mMyh9 embryos, *Myh9*^{tm5RSad} mice are used[24] to breed *Myh9*$^{tm5RSad/tm5RSad}$; *Zp3*$^{Cre/+}$ mothers with WT fathers. To generate mzMyh10 embryos, *Myh10*^{tm7Rsad} mice were used[33] to breed *Myh10*$^{tm7Rsad/tm7Rsad}$; *Zp3*$^{Cre/+}$ mothers with *Myh10*$^{+/-}$ fathers. To generate aPKC-knockout embryos, *Prkci*^{tm1Kido} (ref. 10) and *Prkcz*$^{tm1.1Cda}$ (ref. 4) mice are used to breed *Prkci*$^{-/-}$; *Prkcz*$^{+/-}$ fathers with *Prkci*$^{-/-}$; *Prkcz*$^{tm1.1Cda/+}$; *Zp3*$^{Cre/+}$ mothers.

Mice were used from 6 weeks old onwards.

Chemical reagents. Blebbistatin(+), an inactive enantiomere of the inhibitor, or (−), the selective inhibitor of myosin II ATPase activity (Tocris, 1853 and 1852), and 50 mM DMSO stocks are diluted to 5, 12.5 or 25 μM in KSOM.

Isolation of blastomeres at the 8- and 16-cell stage. Embryos are dissected out of their zona pellucida (ZP) at the 2- to 4-cells stage. ZP-free 8- or 16-cell stage embryos are placed into Ca^{2+}-free KSOM for 5–10 min before being aspirated multiple times (typically between 3 and 5 times) through a narrow glass pipette (with a radius between that of an 8- or 16-cell-stage blastomere and of the whole embryo) until dissociation of cells. To form doublets of polarized and unpolarized 16-cell-stage blastomeres, an 8-cell-stage blastomere is cultured until asymmetric division. To form chimaeras, 16-cell-stage blastomeres are grafted onto a complete embryo using a mouth pipette.

Immunostaining. Primary antibody targeting the double phosphorylated form (Thr18/Ser19) of the myosin regulatory light-chain (Cell Signaling, 3674), PKC-ζ (Santa Cruz, sc-17781), Yap (Abnova, M01, clone 2F12) or its Ser127 phosphorylated form (Cell Signaling, 4911) are used at 1:100. The primary antibody targeting Cdx2 (Biogenex, MU392A-UC) or phosphorylated form (Ser1943) of the non-muscle myosin heavy chain Myh9 (Cell Signaling, 5026) are used at 1:200.

Secondary antibody targeting mouse or rabbit IgG coupled to Alexa Fluor 488 or 546 (Life Technologies) are used at 1:250. Alexa Fluor 633-coupled (ThermoFisher, A22284) phalloidin is used at 1:250.

Micropipette aspiration. As described previously[12], a microforged micropipette coupled to a microfluidic pump (Fluigent, MFCS) is used to measure the surface tension of cells. In brief, micropipettes of radii 7–8 μm for 8-cell-stage embryos and 2.5–3.5 μm for 16-cell-stage embryos are used to apply step-wise increasing pressures on blastomeres until reaching a deformation which has the radius of the micropipette (R_p). At steady state, the surface tension γ_1 of the blastomere is calculated based on Young–Laplace's law: $\gamma_1 = P_c/2(1/R_p - 1/R_c)$, where P_c is the pressure used to deform the cell of radius R_c.

Asymmetric division tracing. To measure the tension of sister cells at the 16-cell stage, time-lapse images of mTmG and H2B–GFP expressing embryos were taken every 30 min from the 8-cell stage onwards. The 8- to 16-cell-stage divisions are tracked, the time lapse is paused to measure the surface tension of both sister cells. After resuming the time lapse, the measured cells are tracked until blastocyst stage. When both sister cells remain at the surface of the embryo, the division is considered symmetric, whereas when one sister cell internalizes during the 16-cell stage and becomes part of the ICM, the division is considered asymmetric.

Microscopy. Tension measurements are performed on a Zeiss Axio Observer microscope with a dry × 20/0.8 PL Apo DICII objective. The microscope is equipped with an incubation chamber to keep the sample at 37 °C. Tension measurements and confocal images are taken using an inverted Zeiss Observer Z1 microscope with a CSU-X1M 5000 spinning disc unit. Excitation is achieved using 488 nm, 561 nm and 633 nm laser lines through a × 63/1.2 C Apo W DIC III water immersion objective. Emission is collected through 525/50 nm, 605/40 nm, 629/62 nm band pass or 640 nm low pass filters onto an EMCCD Evolve 512 camera. The microscope is equipped with an incubation chamber to keep the sample at 37 °C and supply the atmosphere with 5% CO_2.

Data analysis. *Shape analysis.* Using FIJI, we manually fit a circle onto the cell-medium interface to measure the radius of curvature of the cell R_c. We use the angle tool to measure the contact angles θ_1, θ_2 and θ_c. We draw a line perpendicular to the micropipette tip and use the linescan function to measure the diameter of the micropipette and calculate R_p.

Intensity ratio measurements. Using FIJI, we pick confocal slices cutting through the equatorial plane of the apical domain or of two contacting cells. We draw a ~1 μm thick line along the cell-medium interface of the apical and non-apical regions or of each cell and measure the mean intensity. The apical region is defined by visually observing aPKC or mTmG enrichment in the central region of the cell-medium interface. For aPKC-knockout embryos, the actin-rich region at the centre of the cell-medium interface is selected. The transition zones between apical and non-apical or close to cell–cell contacts between polarized and unpolarized cells are excluded to calculate intensity ratios (~5 μm).

Using FIJI, we pick confocal slices cutting through the equatorial plane of the nucleus of a cell. We draw a 2.5 μm radius circle and measure the average intensity in the nucleus and, next to it, in the cytoplasm. We then calculate the nucleus-to-cytoplasm intensity ratio. For whole embryos, the closest distance of the nucleus to the surface of the embryo, marked by phalloidin staining, is measured using the line tool.

Periodic contractions analysis. To analyse periodic contractions, we used a previously described pipeline[12]. In brief, mTmG images are used to segments the cells outlines into 100 equidistant nodes for 8-cell-stage blastomeres and 150 for 16-cell-stage blastomeres doublets. From those nodes, three nodes spaced by 10 or 7 nodes (respectively for 8-cell-stage blastomeres and 16-cell-stage blastomeres doublets) are then taken to fit a circle and compute the local curvature from the inverse radius of this circle. Taking the local curvatures along the cell perimeter over time, a kymograph of local curvature is created. Applying a Fourier transform on the curvature changes over time at each node, we obtain the amplitude of nodes that are classified either as apical or non-apical, based on the mTmG signal that is enriched on the apical domain excluding about 10 nodes around the transition between apical to non-apical domains. For doublets, nodes are classified either as inner or outer cell, based on, when applicable, the asymmetry of the internal contact angles (the cell with the largest internal contact angle being designated as the inner cell for ratio calculation) and/or, when applicable, by the asymmetry in cell size (the smallest cell being designated as the inner cell for ratio calculation). About 10 nodes near the contact edges are excluded from the analysis.

Code availability. Codes are available upon request.

Statistics. Mean, standard deviation, correlation coefficient, two-tailed Student's *t*-test and single-tailed Mann–Whitney *U*-test *P* values are calculated using Excel (Microsoft). Statistical significance of correlation coefficients is obtained from the Pearson correlation table.

The sample size was not predetermined and simply results from the repetition of experiments. No sample was excluded. No randomization method was used. The investigators were not blinded during experiments.

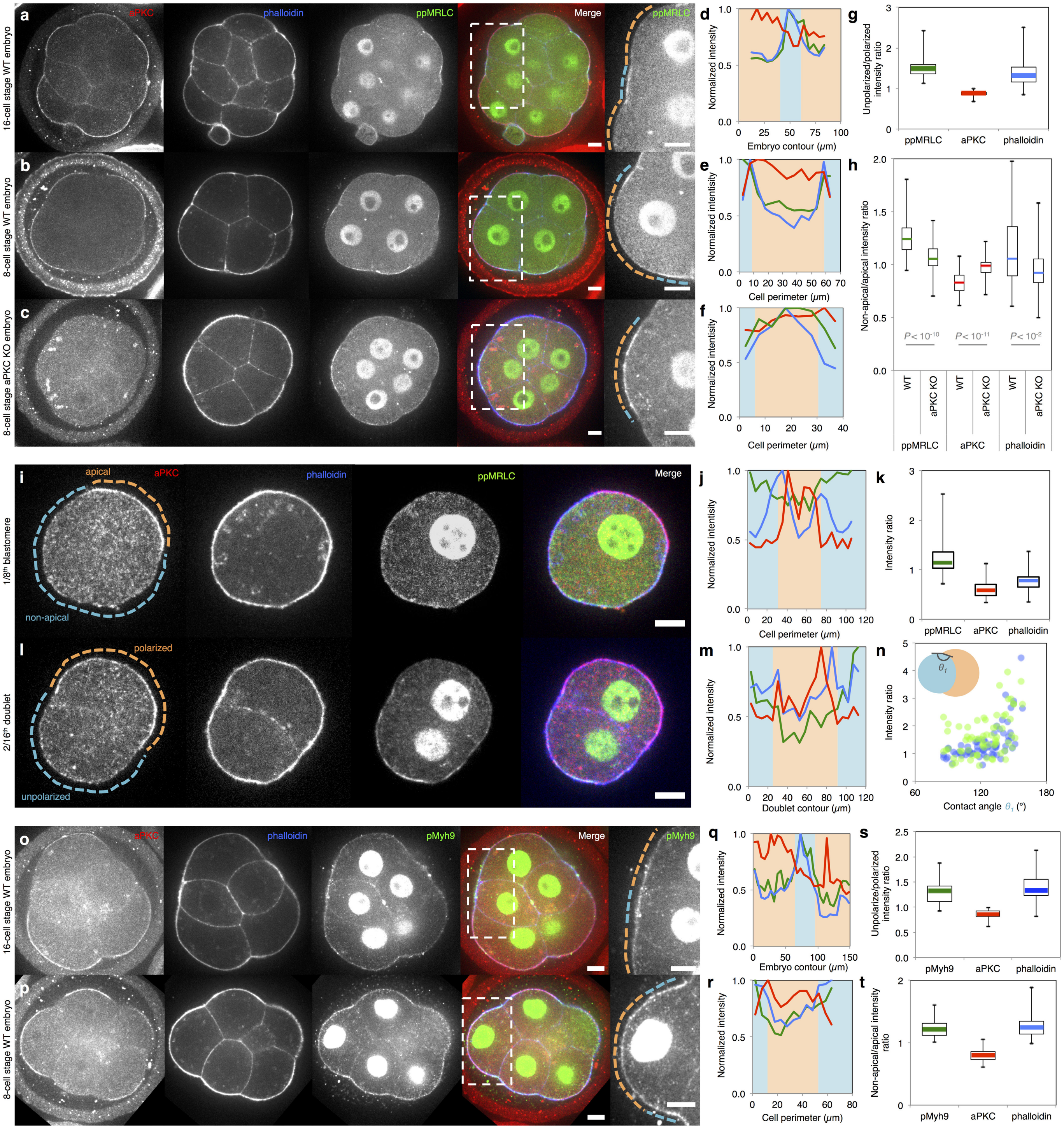
[29] Riedl, J. *et al.* Lifeact mice for studying F-actin dynamics. *Nature Methods* **7,** 168–169 (2010).

[30] Muzumdar, M. D., Tasic, B., Miyamichi, K., Li, L. & Luo, L. A global double-fluorescent Cre reporter mouse. *Genesis* **45,** 593–605 (2007).

[31] Balbach, S. T. *et al.* Nuclear reprogramming: kinetics of cell cycle and metabolic progression as determinants of success. *PLoS ONE* **7,** e35322 (2012).

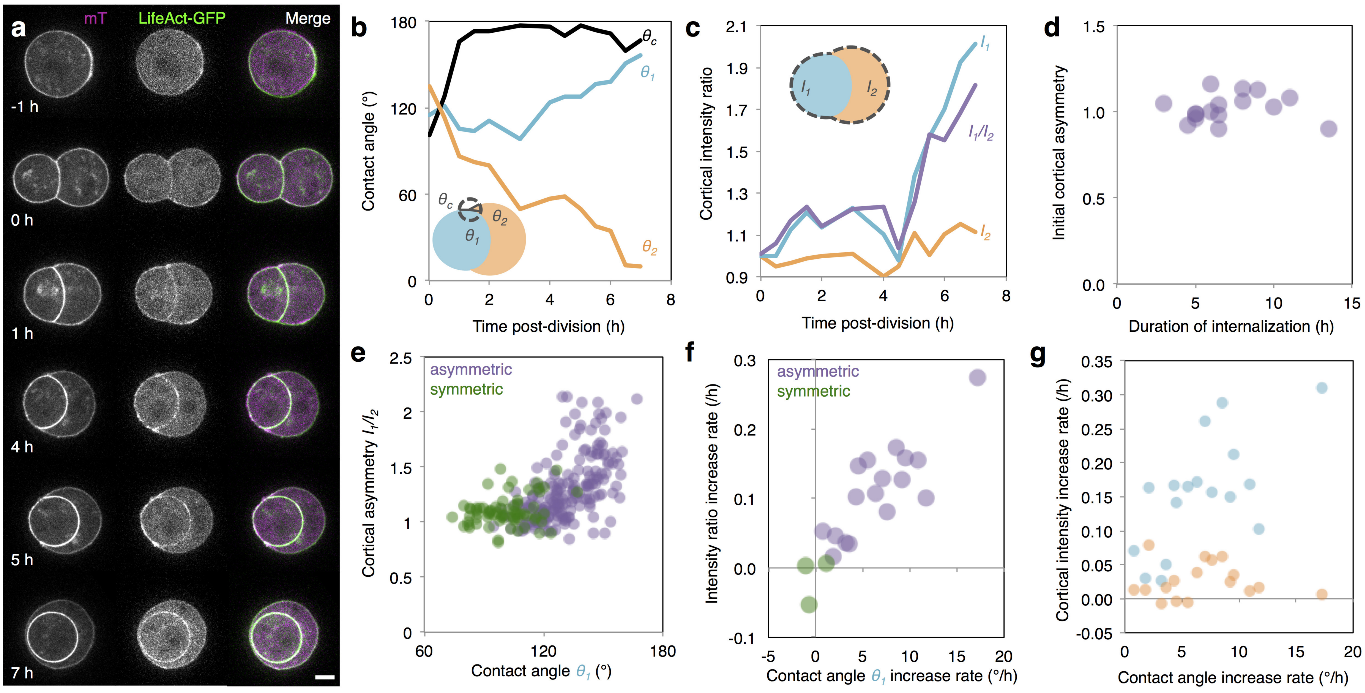
[32] de Vries, W. N. *et al.* Expression of Cre recombinase in mouse oocytes: a means to study maternal effect genes. *Genesis* **26,** 110–112 (2000).

[33] Ma, X. *et al.* Conditional ablation of nonmuscle myosin II-B delineates heart defects in adult mice. *Circ. Res.* **105,** 1102–1109 (2009).

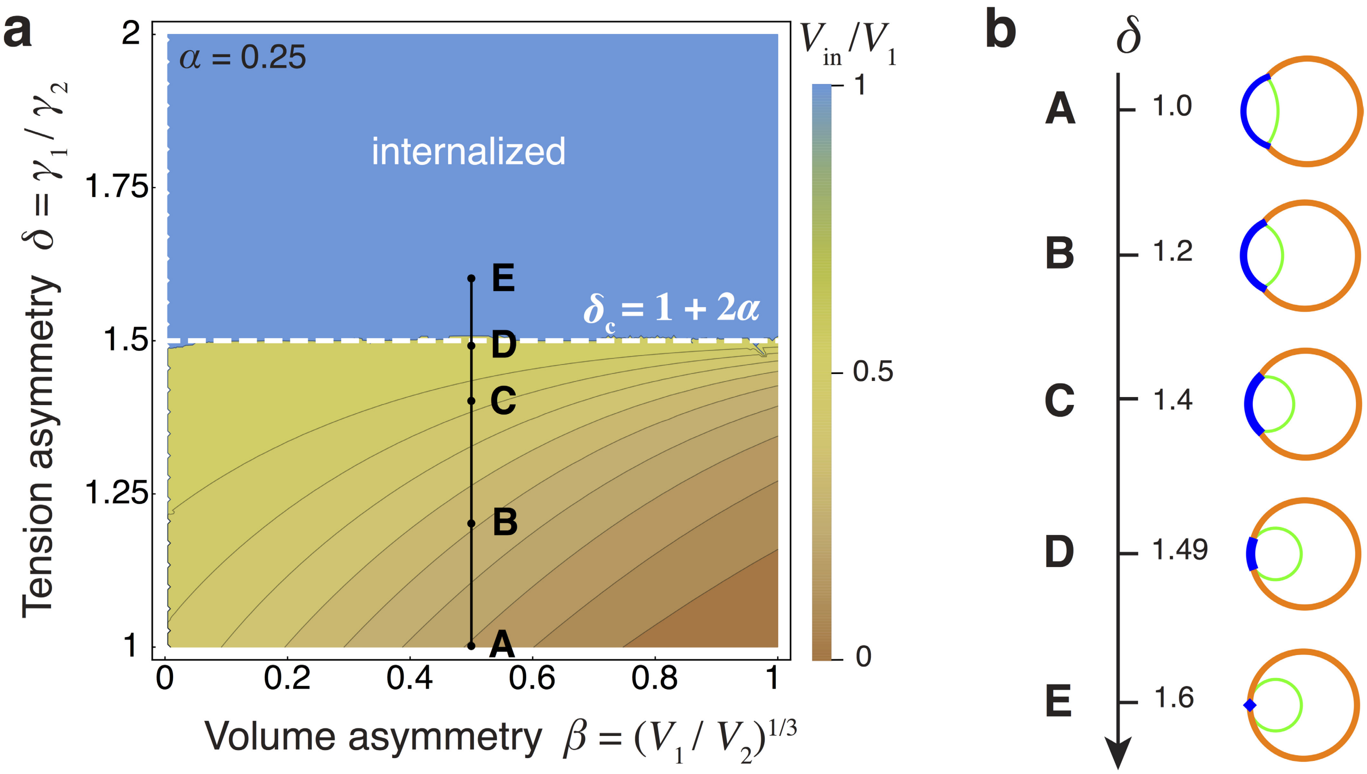


Extended Data Figure 1 | See next page for caption.

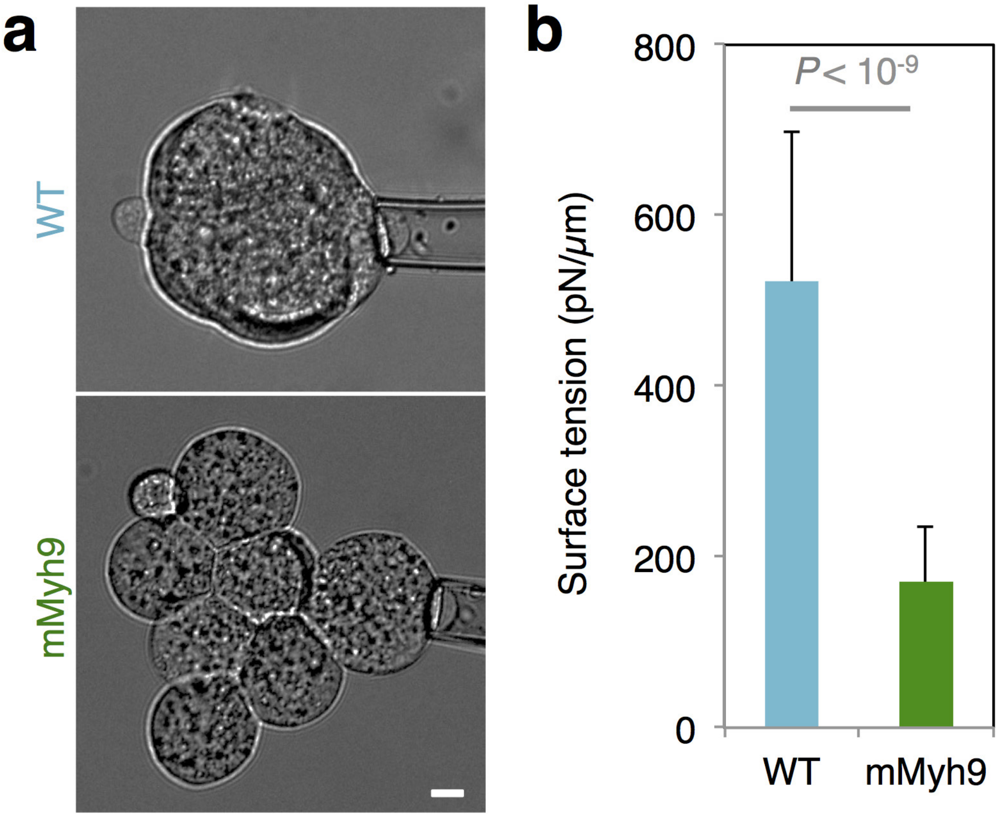
Extended Data Figure 1 | aPKC antagonizes myosin phosphorylation at the apical domain. a–**c**, Immunostaining of 16- (**a**) and 8-cell-stage wild-type (**b**) and aPKC-knockout (**c**) embryos showing aPKC (red), phalloidin (blue) and bi-phosphorylated myosin regulatory light chain (ppMRLC; green). Enlarged images of ppMRLC are shown on the far right. **d**–**f**, Cortical intensity profiles under the dotted lines in the far right panels of **a**–**c**. Apical domains are highlighted in orange and non-apical regions in blue. **g**, **h**, Box plot of unpolarized/polarized blastomere intensity ratio at the 16-cell stage (43 neighbouring blastomeres from 35 embryos from 3 experiments) and non-apical/apical intensity ratio at the 8-cell stage for wild-type (WT) and aPKC-knockout (KO) embryos (68 and 58 blastomeres from 22 and 12 embryos from 2 and 3 experiments, respectively). ppMRLC is in green, aPKC in red and phalloidin in blue. Student's *t*-test *P* values between wild type and aPKC knockout. At the 16-cell stage (**a**, **d**, **g**), blastomeres showing accumulations of ppMRLC and phalloidin at their cell-medium interfaces have less aPKC. At the 8-cell stage, the apical domain does not occupy the entirety of the cell-medium interface (**b**, **e**). The aPKC-rich apical domain shows less myosin than the aPKC-poor region of the cortex (**b**, **e**, **h**). In aPKC-knockout embryos (**c**, **f**, **h**), no aPKC-rich, nor ppMRLC-poor regions can be observed at the cell-medium interface of blastomeres. **i**, Immunostaining of 8-cell-stage blastomeres showing aPKC (red), phalloidin (blue), ppMRLC (green) and merged staining. The apical domain is highlighted in orange, the non-apical cortex in blue. Scale bar, 10 μm. **j**, Intensity profile along the cell perimeter showing aPKC (red), phalloidin (blue) and ppMRLC (green). The apical intensity is highlighted in orange, the non-apical cortex in blue. **k**, Box plot of apical and non-apical intensity ratio of cortical aPKC (red), phalloidin (blue) and ppMRLC (green) for 27 blastomeres from 3 experiments. Blastomeres isolated at the 8-cell stage show an aPKC-rich region (**i**–**k**), which has less cortical ppMRLC and phalloidin than the aPKC-poor region of the cell-medium interface. The ppMRLC- and actin-rich regions are distinct from the basolateral domain of cells (their cell–cell contact), which have less ppMRLC and actin (**a**–**c**, **l**)[12]. This ppMRLC cortical region is therefore labelled 'non-apical'. **l**, Immunostaining of doublets of 16-cell-stage blastomeres showing aPKC (red), phalloidin (blue), ppMRLC (green) and merged staining. The polarized blastomere is highlighted in orange, the unpolarized one in blue. Scale bar, 10 μm. **m**, Intensity profile along the doublet perimeter showing aPKC (red), phalloidin (blue) and ppMRLC (green). The polarized blastomere is highlighted in orange, the unpolarized one in blue. Blastomeres isolated at the 8-cell stage divide to give rise to doublets of 16-cell-stage blastomeres[2,16]. The polarized sister cell shows high aPKC and low ppMRCL/phalloidin at their cell-medium interfaces when compared to the non-polarized sister cell (**l**, **m**). **n**, Cortical intensity ratio of ppMRLC (green) and phalloidin (blue) between the inner and outer cells as a function of the inner contact angles θ_1 (Pearson $R = 0.464$ and 0.614, $n = 67$ doublets from 2 experiments, $P < 0.001$). During the 16-cell stage, polarized blastomeres can envelop their unpolarized sister blastomeres (Supplementary Video 5)[2,16]. As envelopment occurs, the internal contact angles change (Extended Data Fig. 2). As the internal contact angles change, the asymmetry in cortical ppMRLC and phalloidin between sister blastomeres changes. After another division, a cyst consisting of four blastomeres forms (Supplementary Video 5). This structure is equivalent to the blastocyst in terms of gene expression[16] (Fig. 4). **o**, **p**, Immunostaining of 16- (**o**) and 8-cell-stage (**p**) embryos showing aPKC (red), phalloidin (blue) and myosin heavy chain phosphorylated on S1943 (pMyh9; green). Enlarged images of ppMRLC are shown on the far right. **q**, **r**, Cortical intensity profiles under the dotted lines on the far right of **o**, **p**. Apical domains are highlighted in orange and non-apical regions in blue. **s**, **t**, Box plot of unpolarized/polarized blastomere intensity ratio at the 16-cell stage (24 neighbouring blastomeres from 16 embryos from 3 experiments) and non-apical/apical intensity ratio at the 8-cell stage (34 blastomeres from 10 embryos from 3 experiments). pMyh9 in green, aPKC in red and phalloidin in blue.



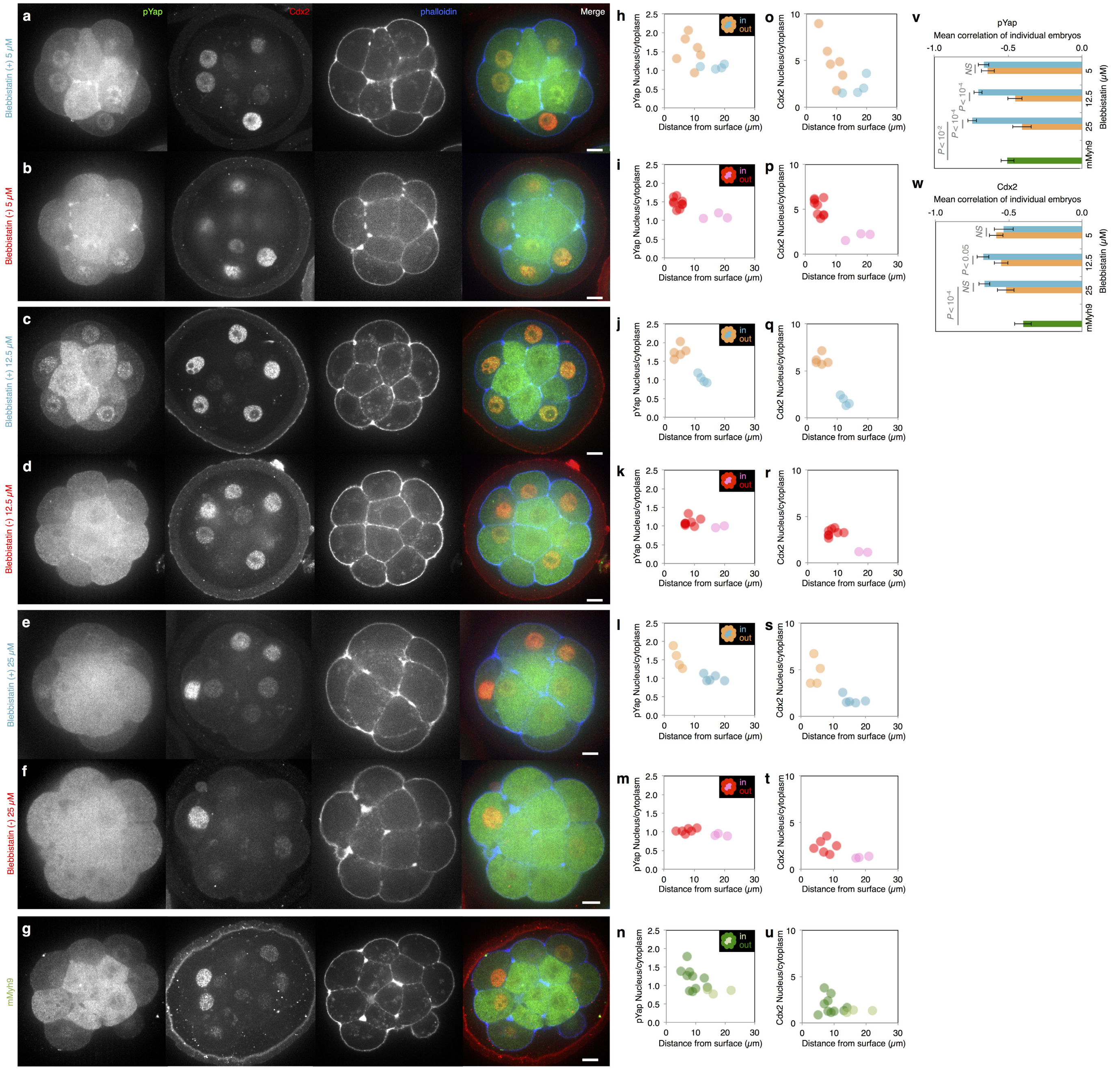
Extended Data Figure 2 | Cortical asymmetries intensify during the 16-cell stage. a, Time-lapse of mTmG (magenta) and LifeAct–GFP (green) expressing doublets of 16-cell-stage blastomeres. Scale bar, 10 µm. **b**, **c**, External (θ_c), and internal (θ_1 and θ_2) contact angles (**b**) and cortical LifeAct–GFP intensities of unpolarized I_1 and polarized I_2 blastomeres and intensity ratio I_1/I_2 (**c**) over time for the doublet shown in **a**. Blastomeres isolated at the 8-cell stage can divide asymmetrically to give rise to a polarized blastomere that will envelop its unpolarized sister blastomere (**a**). The external contact angle θ_c shows a rapid re-compaction of the cell doublet after division (**b**). The internal contact angles θ_1 and θ_2 indicate the progression of the envelopment process (**b**). As this happens, the cortical intensity of LifeAct–GFP of the internalizing blastomere I_1 increases while the one of the enveloping blastomere I_2 remains comparably more stable (**c**). This increases the cortical asymmetry I_1/I_2 (**c**). **d**, Initial cortical asymmetry over internalization time of doublets of 16-cell-stage blastomeres (Pearson $R = 0.064$, $n = 16$ doublets from 4 experiments, $P > 0.1$). The initial cortical asymmetry, calculated within 30 min after division, is 1.0 ± 0.1 (mean ± s.d., $n = 17$ doublets from 4 experiments) and does not control the time it takes for envelopment to occur (**d**). **e**, Intensity ratio as a function of the contact angle θ_1 of doublets throughout the 16-cell-stage blastomeres (Pearson $R = 0.573$, 186 measurements on 17 asymmetric doublets (purple), $P < 0.001$ and Pearson $R = 0.266$, 69 measurements on 3 symmetric (green) doublets, $P < 0.1$, from 4 experiments). As the internal contact angles change, the asymmetry in cortical LifeAct–GFP between sister blastomeres with distinct polarity changes. **f**, Cortical intensity ratio increase rate as a function of the contact angle θ_1 increase rate (Pearson $R = 0.824$, $n = 17$ asymmetric (purple), $P < 0.001$, and Pearson $R = 0.393$, $n = 3$ symmetric (green) doublets, $P > 0.1$, from 4 experiments). **g**, Cortical intensity increase rate as a function of the contact angle increase rate for the polarized (orange, Pearson $R = -0.026$, $n = 17$ asymmetric doublets, $P > 0.1$) and unpolarized blastomere (blue, Pearson $R = 0.658$, $n = 17$ asymmetric doublets, $P < 0.01$) of a doublet resulting from asymmetric division or of two polarized cells resulting from a symmetric division (green, Pearson $R = -0.011$, $n = 3$ symmetric doublets, $P > 0.1$), from 4 experiments. The rates are correlated, which suggests that the dynamics of internalization and the dynamics of building up of cortical asymmetries are linked.



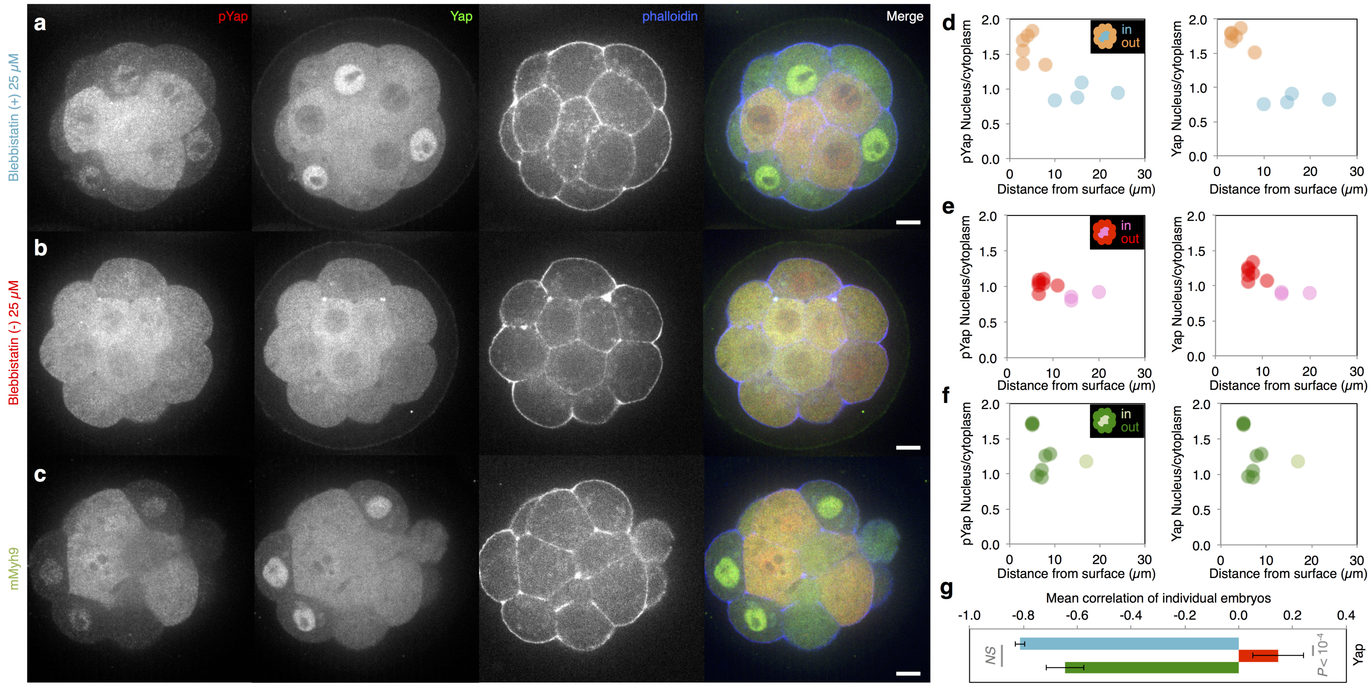
Extended Data Figure 3 | Cell size has no influence on internalization. Phase diagram describing the mechanical equilibrium of a cell within a doublet or embryo as function of the cell size asymmetry parameter β and the tension asymmetry parameter δ, for a fixed compaction parameter $\alpha = 0.25$. The colour code measures the degree of internalization, defined as the proportion of internalized volume V_{in}/V_1, which equals 1 for the internalized cell. The dotted line indicates the threshold value δ_c at which internalization occurs. An example of internalization with $\beta = 0.5$ is indicated in black (from A to E). Changing the volume asymmetry does not change the internalization threshold. Internalization of a doublet with $\beta = 0.5$ obtained with the analytical model for the same values of δ as indicated in the diagram from A to E.



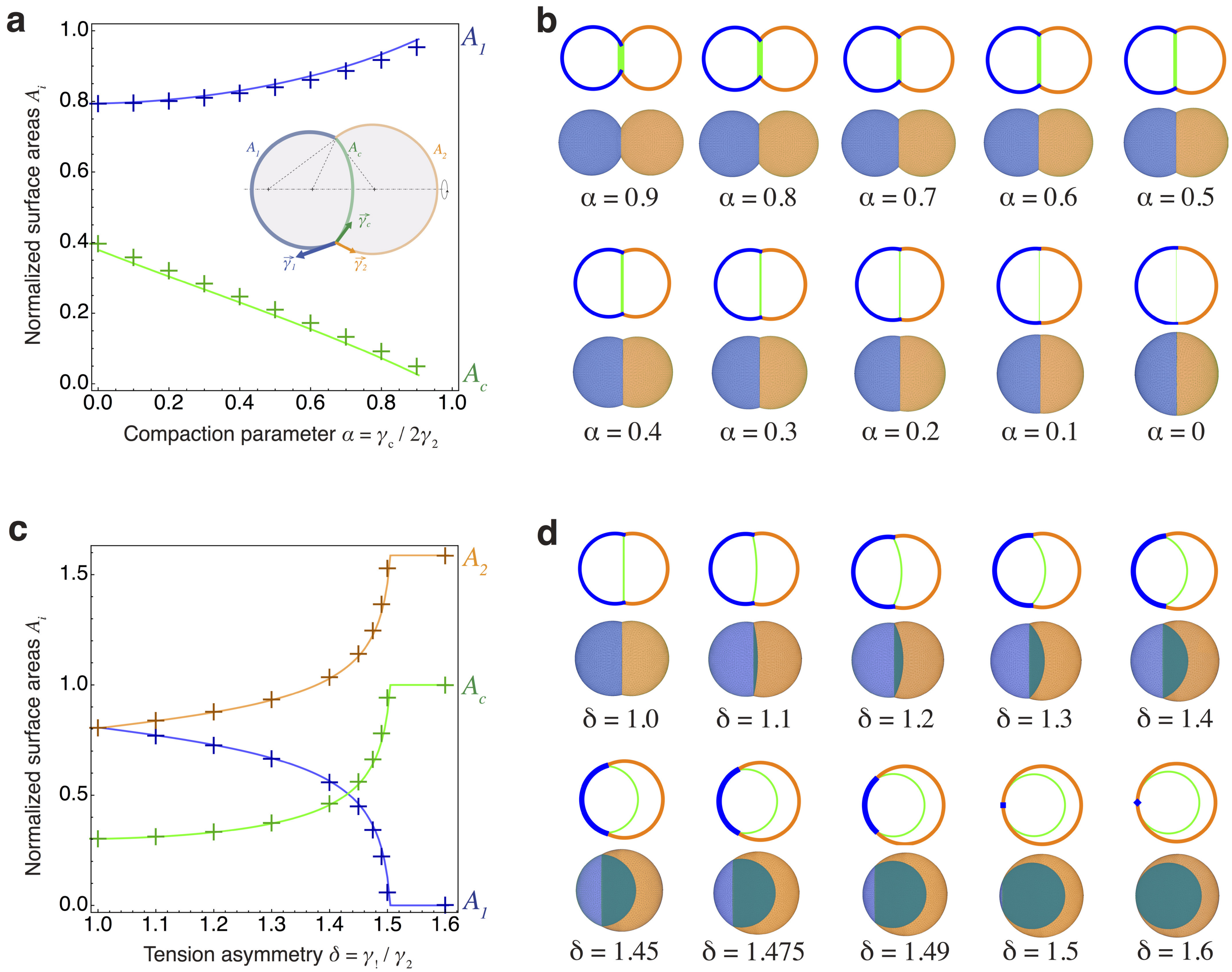
Extended Data Figure 4 | Contractility is required for internalization. Brightfield images of tension measurement on wild-type (WT; top) and mMyh9 (bottom) 8-cell-stage embryos. Scale bar, 10 μm. Mean ± s.d. of 25 blastomeres from 4 wild-type embryos and 26 blastomeres from 7 mMyh9 embryos from 2 experiments, Student's *t*-test $P < 10^{-9}$.



Extended Data Figure 5 | Control of pYap and Cdx2 localization by contractility in a dose-dependent manner. a–g, Immunostaining of wild-type embryos treated for 3 h with Bb(+) at 5 (**a**), 12.5 (**c**) or 25 (**e**) µM or with Bb(−) at 5 (**b**), 12.5 (**d**) or 25 (**f**) µM or of mMyh9 embryos (**g**) showing pYap (green), Cdx2 (red) and phalloidin (blue). **h–u,** Nucleus-to-cytoplasm intensity ratio of pYap (**h–n**) or Cdx2 (**o–u**) as a function of the distance from the surface for wild-type embryo treated with Bb(+) (outer cells in orange and inner cells in blue) at 5 (**h**, **o**, corresponding embryo shown in **a**), 12.5 (**j**, **q**, corresponding embryo shown in **c**) or 25 (**l**, **s**, corresponding embryo shown in **e**) µM or with Bb(−) (outer cells in red and inner cells in pink) at 5 (**i**, **p**, corresponding embryo shown in **b**), 12.5 (**k**, **r**, corresponding embryo shown in **d**) or 25 (**m**, **t**, corresponding embryo shown in **f**) µM and for mMyh9 embryos (**n**, **u**, corresponding embryo shown in **g**). **v**, **w**, Mean ± s.e.m. Pearson correlation values between the nucleus to cytoplasm intensity ratio of pYap (**v**) or Cdx2 (**w**) as a function of the distance from the surface from individual embryos. Two-hundred and seven blastomeres from 20 embryos for Bb(+) 5 µM, 252 blastomeres from 29 embryos for Bb(+) 12.5 µM, 179 blastomeres from 18 embryos for Bb(−) 5 µM and 267 blastomeres from 28 embryos for Bb(−) 12.5 µM from 3 experiments each. Two-hundred and eighty-one cells from 28 embryos from 5 experiments for pYap and 136 cells from 13 embryos from 4 experiments for Cdx2 for 25 µM Bb(+), 241 cells from 32 embryos from 5 experiments for pYap and 192 cells from 22 embryos from 4 experiments for Cdx2 for 25 µM Bb(−), and 349 cells from 32 embryos from 6 experiments for pYap and 217 cells from 21 embryos from 3 experiments for Cdx2 for mMyh9. Student's *t*-test *P* values; NS, not significant.



Extended Data Figure 6 | Contractility controls Yap subcellular localization. a–c, Immunostaining of wild-type embryos treated with 25 μM Bb(+) (**a**; an inactive enantiomere of the inhibitor) or Bb(−) (**b**; the selective inhibitor of myosin II ATPase activity) for 3 h or mMyh9 embryos (**c**) showing Yap (green), pYap (red) and phalloidin (blue). **d–f**, Nucleus-to-cytoplasm intensity ratio of pYap (left) and Yap (right) as a function of the distance from the surface for wild-type embryo treated with 25 μM Bb(+) (**d**; outer cells in orange and inner cells in blue, corresponding embryo shown in **a**) or Bb(−) (**e**; outer cells in magenta and inner cells in red, corresponding embryo shown in **b**) or mMyh9 embryo (**f**; outer cells in dark green and inner cells in light green, corresponding embryo shown in **c**). **g**, Mean ± s.e.m. Pearson correlation values between the nucleus to cytoplasm intensity ratio of Yap as a function of the distance from the surface from individual embryos. Two-hundred and fifty-two cells from 29 embryos for Bb(+), 201 cells from 26 embryos for Bb(−) and 132 cells from 12 embryos for mMyh9 from 3 experiments each. Student's *t*-test *P* value is shown; NS, not significant.



Extended Data Figure 7 | Quantitative comparison between analytical and numerical results. **a**, Comparison of the surface areas of the cell medium (blue) and cell–cell interfaces (green) between the simulations (crosses) and the analytical model (lines) for different values of the compaction parameter α between 0 and 1. A schematic diagram of a cell doublet defining the cell medium and cell–cell surface tensions γ_1, γ_2 and γ_c and areas A_1, A_2 and A_c are shown as an inset. **b**, Configurations of doublets as predicted by the analytical model and simulations for the discrete values of α corresponding to the plot in **a**. **c**, Comparison of the surface areas of the cell-medium interfaces of cell 1 (blue), 2 (orange) and of the cell–cell interface (green) between the simulations (crosses) and the analytical model (lines) for different values of the tension asymmetry parameter δ between 1 and 1.6 fixed compaction parameter $\alpha = 0.25$. **d**, Configurations of doublets as predicted by the analytical model and simulations for the discrete values of δ corresponding to the plot in **c**.

LETTER

doi:10.1038/nature19098

HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis

David A. Jacques[1], William A. McEwan[1], Laura Hilditch[2], Amanda J. Price[1]†, Greg J. Towers[2] & Leo C. James[1]

During the early stages of infection, the HIV-1 capsid protects viral components from cytosolic sensors and nucleases such as cGAS and TREX, respectively, while allowing access to nucleotides for efficient reverse transcription[1]. Here we show that each capsid hexamer has a size-selective pore bound by a ring of six arginine residues and a 'molecular iris' formed by the amino-terminal β-hairpin. The arginine ring creates a strongly positively charged channel that recruits the four nucleotides with on-rates that approach diffusion limits. Progressive removal of pore arginines results in a dose-dependent and concomitant decrease in nucleotide affinity, reverse transcription and infectivity. This positively charged channel is universally conserved in lentiviral capsids despite the fact that it is strongly destabilizing without nucleotides to counteract charge repulsion. We also describe a channel inhibitor, hexacarboxybenzene, which competes for nucleotide binding and efficiently blocks encapsidated reverse transcription, demonstrating the tractability of the pore as a novel drug target.

There is increasing evidence that the HIV-1 capsid remains intact as it traverses the cytoplasm of a newly infected cell. Prematurely uncoated viruses trigger innate immune sensing[2], assembled capsid proteins are required to properly engage the nuclear pore complex[3], and intact capsids have been observed at the nuclear envelope[4]. Reverse transcription has been postulated to occur within the HIV-1 virion during cytoplasmic transit, yet structural analyses of the HIV-1 capsid have not defined a pore through which small molecules such as deoxynucleoside triphosphates (dNTPs) might pass. One possible location for a pore would be the six-fold axis at the centre of each capsid protein (CA) hexamer, but this is not evident from existing hexamer structures, as it is obscured by the N-terminal β-hairpin. By comparing all of the available CA crystal structures with resolved β-hairpins[5–11], including monomeric crystal forms, we observed that the β-hairpin can adopt alternate conformations that differ by up to 15 Å (as measured by the displacement of Q7) (Fig. 1a). When reconstructed in the context of a hexamer, several of these β-hairpin conformations result in a pore about the six-fold axis (Fig. 1b). The different β-hairpin conformations are the result of a pivoting movement of up to 37.5 ° about the N-terminal proline, an essential capsid residue that forms a salt-bridge with D51 (ref. 12) (Fig. 1a, Supplementary Video 1). In structures where the pore would be open, D51 also participates in a second salt-bridge interaction with H12. Conversely, in structures where the pore would be closed, including all previously solved disulfide-stabilized hexamers ($CA_{hexamer}$), a water molecule has displaced the H12 side chain and coordinates a tetrahedral hydrogen-bond network between H12, T48, Q50 and D51. We hypothesized that the protonation state of H12 may

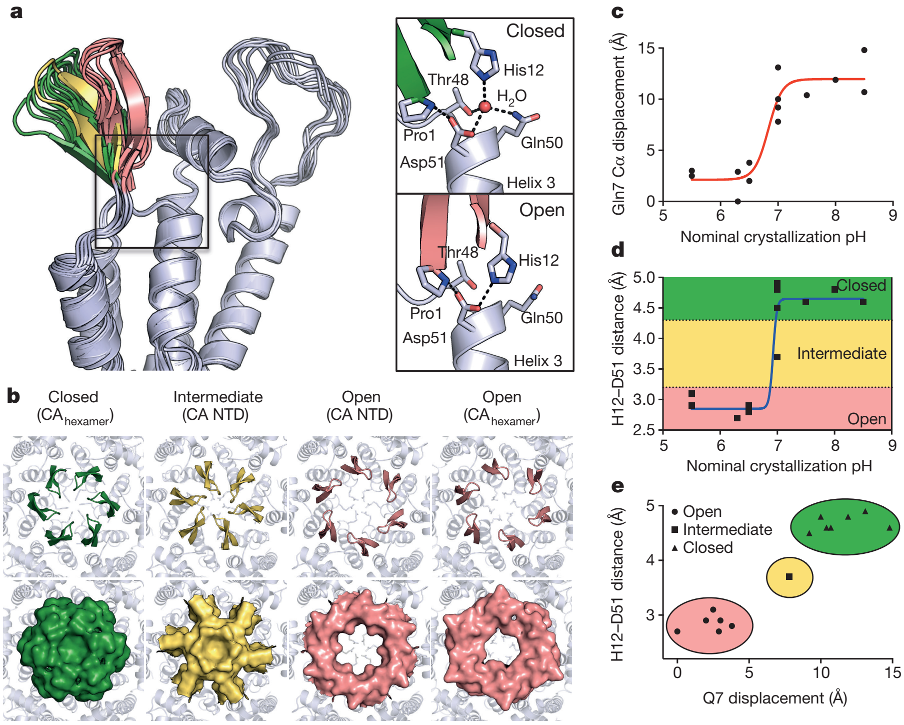


Figure 1 | HIV-1 capsid hexamers have a pore at the six-fold symmetry axis. a, Superposition of N-terminal domains from solved capsid structures. A detailed view of the boxed region shows that the β-hairpin toggles between closed (green) and open (pink) states as a result of the hydrogen-bond network around P1, H12 and D51. **b**, β-hairpin (coloured) conformations dictate the presence of a pore at the six-fold axis. Hexamers of CA N-terminal domain (CA NTD) structures have been assembled using symmetry operators from $CA_{hexamer}$ structures. **c**, **d** Displacement of Q7 (**c**) and H12–D51 distance (**d**) as a function of crystallization pH. **e**, Correlation of Q7 displacement with H12–D51 distance.

[1]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge Biomedical Campus, Cambridge CB2 0QH, UK. [2]Infection and Immunity, University College London, Cruciform Building 3.3, 90 Gower Street, London WC1E 6BT, UK. †Present address: Astex Pharmaceuticals, 436 Cambridge Science Park, Milton Road, Cambridge, CB4 0QA, UK.

be crucial in determining which arrangement is favoured and therefore that the conformation of the β-hairpin in published structures will have been influenced by the pH at which they were solved. Notably, when the relative displacement of the β-hairpin (Q7 Cα) is plotted against crystallization pH, the structures resolve into two groups; at pH <7 an open-pore β-hairpin conformation is observed, whereas at pH >7, a closed-pore conformation is favoured (Fig. 1c, Extended Data Table 1). The same correlation is observed when the distance between D51 and H12 is plotted against pH (Fig. 1d, e), confirming the importance of H12 in determining β-hairpin conformation. Structures solved at pH 7 display the greatest β-hairpin variability, consistent with maximum pore flexibility occurring under physiological conditions. The likely reason why a pore has not been detected in published $CA_{hexamer}$ structures is because they were solved at a basic pH where H12 is deprotonated and a closed pore is favoured. To test this hypothesis and demonstrate that the pore can open in the context of an assembled hexamer, we sought to crystallize $CA_{hexamer}$ under acidic conditions. We obtained a previously unreported crystal form at pH 5.5, the structure of which contains a β-hairpin in the open conformation and an exposed pore (Fig. 1b, far right).

Using all available CA structures to define a range of movement for the β-hairpin, we observed that the pivoting about P1 results in an iris-like motion, which creates an aperture on the outer surface of the capsid (Fig. 2a). In the open state, a chamber that is 25 Å deep and 3,240 $Å^3$ in volume is revealed, which culminates in a ring of six arginine side chains from residue 18 (Fig. 2b, Supplementary Video 2). This cluster of basic residues in close proximity results in highly electropositive foci at the centre of each hexamer. The R18 residues adopt multiple conformations (Extended Data Fig. 1a) to give a maximum pore diameter of 8 Å, sufficient to allow transit of a dNTP molecule. We therefore reasoned that this feature might provide an efficient means to recruit dNTPs into the capsid interior while excluding larger molecules. We therefore tested whether $CA_{hexamer}$ can interact with dNTPs by fluorescence anisotropy and found that all four nucleotides bind with a remarkably high affinity of between 6–40 nM (Fig. 2c). All biophysical measurements were undertaken in an 'intracellular buffer' (see Methods) that is designed to match salt concentrations in the cell. We also observed that physiological concentrations of inorganic phosphate had little effect on dNTP binding and that the pore could not distinguish between dNTPs and ribonucleoside triphosphates (rNTPs) (Extended Data Fig. 2) consistent with the observation that ribonucleoside monophosphates are often incorporated into newly synthesized viral DNA[13]. Analysing the kinetics of interaction by stopped-flow revealed that binding is driven by an extremely rapid on-rate of $>2 \times 10^8\,M^{-1}\,s^{-1}$, although this is probably an underestimate as the reaction becomes immeasurably fast at increasing reactant concentrations (Fig. 2d, e). Separate dissociation experiments in which fluorescent deoxycytidine triphosphate (dCTP) was displaced with excess unlabelled dCTP determined that the off-rate is also fast at $>12\,s^{-1}$ (Fig. 2f), equivalent to a half-life of 58 ms. Calculation of on-rates for all four nucleotides on the basis of their steady-state affinities and off-rates confirms that HIV-1 hexamers achieve association rates between 10^8–$10^9\,M^{-1}\,s^{-1}$. These are unusually rapid association kinetics, typically found in enzymes that have achieved so-called kinetic perfection. Such ultra-rapid enzymes are rare because of the strong fitness advantage needed for their selection over merely very fast equivalents[14]. The rapid on-rate of dNTP recruitment that HIV achieves may be the result of an electrostatically assisted association binding mechanism, as has been described for the barnase–barstar complex[15]. Importantly, the combination of fast on and off rates suggests that although the HIV-1 capsid may recruit dNTPs efficiently, these nucleotides quickly dissociate to become available as substrates for reverse transcription.

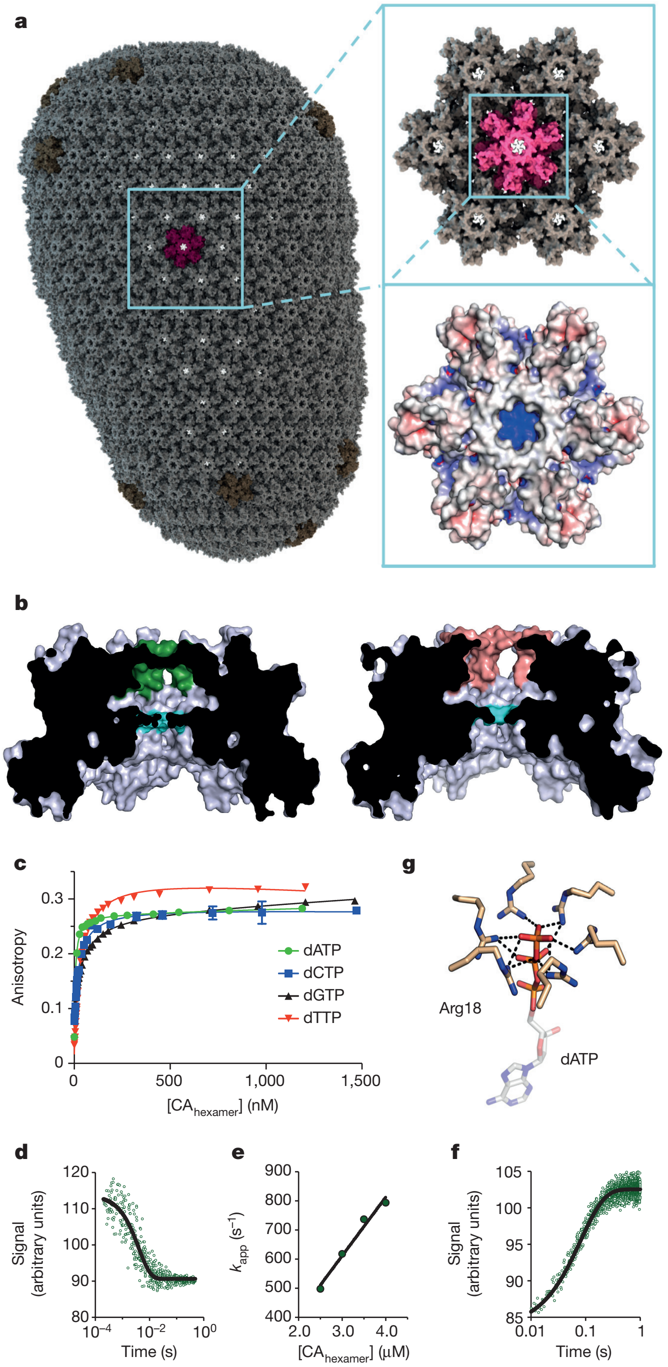


Figure 2 | The HIV-1 capsid pore is strongly electropositive and recruits dNTPs with rapid association and dissociation kinetics. a, Model of an HIV-1 virion with hexamers in an open conformation reveals that the capsid is porous. Surface electrostatic potential shows that the pores are highly electropositive. **b**, Cross sections through the closed (β-hairpin green) and open (β-hairpin pink) $CA_{hexamer}$ showing a central chamber that is accessible in the open state. R18 (cyan) creates a bottleneck at the base of the chamber underneath the β-hairpin. **c**, Fluorescence anisotropy measurements of dNTPs binding to $CA_{hexamer}$ (mean of quadruplicate measurements ± s.d.). **d**, Example of pre-steady-state association kinetics of dCTP with $CA_{hexamer}$. **e**, Apparent rate constant (k_{app}) at increasing $CA_{hexamer}$ concentrations. **f**, Dissociation of unlabelled dCTP:$CA_{hexamer}$ by excess fluorescent dCTP. **g**, R18 coordinates the phosphates in a dATP-bound $CA_{hexamer}$ structure.

To test whether the ring of arginine residues is responsible for nucleotide recruitment, we solved the structure of HIV-1 $CA_{hexamer}$ in complex with dATP and found that it binds as predicted in the centre of the arginine ring via its phosphate groups (Fig. 2g). Although there is electron density for the phosphates, the position of the base can only be modelled, probably because hexamer rotational symmetry

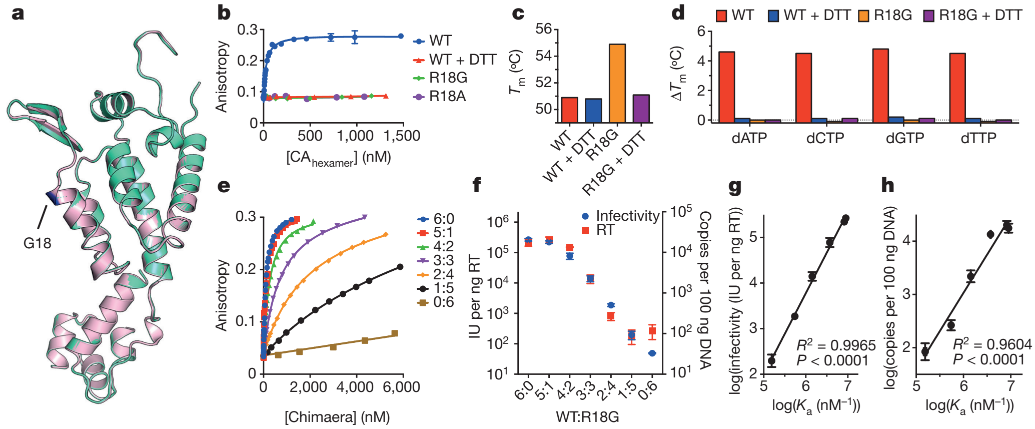


Figure 3 | R18 is crucial for nucleotide recruitment, reverse transcription and infectivity. a, Superposed monomers of R18G (light pink) and wild-type (light green) $CA_{hexamer}$. **b**, Binding of capsid variants to dCTP as measured by fluorescence anisotropy. **c**, DSF stability measurements expressed as T_m for wild type and R18G ± DTT. **d**, DSF measurements of the effect of dNTPs on the stability of wild type and R18G expressed as ΔT_m relative to unbound. **e**, Fluorescence anisotropy titrations of dTTP binding by chimaeric $CA_{hexamers}$ with different R:G ratios at position 18. **f**, Comparison of infectivity and reverse transcription (RT) of chimaeric viruses. IU, infectious units. **g**, **h**, Correlation between HIV-1 capsid dTTP affinity, viral infectivity (**g**) and reverse transcription (**h**). Anisotropy measurements are mean of quadruplicate ± s.d., while infectivity and RT are mean of triplicate ± s.d.

allows the dATP base to occupy six equivalent positions, averaging its density over a large volume (Extended Data Fig. 1). Comparison with a structure solved under identical conditions, but in the absence of dATP, confirms that the observed density corresponds to the nucleotide. To investigate further the importance of R18 in the recruitment of dNTPs, we produced R18G and R18A $CA_{hexamer}$ mutants. Neither R18G nor R18A affected the overall structure of the protein and neither displayed measurable nucleotide binding (Fig. 3a, b). To determine whether formation of the arginines into a ring is required, we performed binding experiments on wild-type protein in the presence of dithiothreitol (DTT), which reduces the disulfide bonds that stabilize the hexameric construct, resulting in monomeric CA. No binding was observed to monomeric CA, demonstrating that once the pore is disassembled, the capsid can no longer recruit dNTPs (Fig. 3b).

The concentration of positive charge provided by the R18 ring is an unusual feature and might be expected to exert a destabilizing influence on the capsid lattice. Conversely, it has been calculated that arginine pairs can stabilize protein interfaces[16] and arginine clusters have been postulated to have a stabilizing effect[17]. We performed differential scanning fluorimetry (DSF) to compare the relative stability of $CA_{hexamer}$ with and without the electropositive pore. We observed remarkable stability of the R18G hexamer relative to the wild-type complex, corresponding to an unexpectedly large increase in melting temperature (T_m) of 4 °C (Fig. 3c, Extended Data Fig. 3). A similar increase in stability was observed in the wild-type hexamer in the presence of dNTPs, whereas no stabilization was observed when dNTPs were added to R18G hexamer (Fig. 3d). Taken together, these results suggest that the pore is indeed a destabilizing feature that is tolerated by the capsid lattice in order to facilitate nucleotide binding. An alignment of capsid sequences predicts an electropositive pore to be conserved across retrovirus genera, with the exception of the gammaretroviruses (Extended Data Fig. 4). Although there appears to be no R18 equivalent in gammaretroviruses, analysis of the published murine leukaemia virus (MLV) capsid structure reveals a large channel running down the six-fold axis with an inward facing Arg residue at position 3 (ref. 18). This residue may have a role in attracting dNTPs, but it is unlikely that the MLV capsid has the same size selectivity as the HIV capsid as the pore is much wider.

The observation that HIV-1 has evolved the fastest possible rate constant for nucleotide recruitment suggests that dNTP import may be a limiting factor in reverse transcription and infectivity. To determine how the efficiency of nucleotide recruitment affects these measures of viral fitness, we constructed a matched set of chimaeric wild-type:R18G $CA_{hexamer}$ and viruses (see Extended Data Fig. 5 for chimaera controls). R18G was chosen as its capsid morphology has previously been demonstrated to be indistinguishable from wild type[19]. Furthermore, R18G was able to saturate the activity of capsid-binding restriction factor TRIM5α, confirming that assembled capsids enter the cytoplasm[20] (Extended Data Fig. 6a). We tested our chimaeric hexamers for dNTP binding and found that an incorporation ratio of 5:1 (five arginines to one glycine) had a minimal effect on affinity (Fig. 3e). However, as the proportion of glycine residues was increased, there was a dose-dependent decrease in dNTP binding. Testing HIV-1 green fluorescent protein (GFP) vesicular stomatitis virus (VSV-G) pseudotyped chimaeric viruses for infectivity revealed a similar pattern of R18 dependence, in which there was little change in infectivity at a ratio of 5:1 but a dominant negative effect at higher G18 ratios (Fig. 3f). Assuming a binomial distribution of arginines and glycines in viral hexamers, the data fit a model in which removal of two or more arginines from the pore is detrimental to the virus (Extended Data Fig. 6b, c), which is consistent with the observation that removal of one arginine has little effect on dNTP affinity. Importantly, there is close correlation between chimaera infectivity and nucleotide affinity, consistent with the recruitment of dNTPs impacting directly on viral infection (Fig. 3g). Such a mechanism would be expected to influence infection at the level of reverse transcription, and indeed a similarly close correlation is observed between chimaera affinity and the production of early reverse transcripts (Fig. 3h).

We propose that reverse transcription takes place within the protected environment of the capsid by recruiting nucleotides through a strongly electropositive pore at the centre of each capsid hexamer. In order to explore this further we performed endogenous reverse transcription (ERT) assays in which HIV-1 capsid cores were purified from virions[21] and their reverse transcriptase activity quantified *in vitro* (Extended Data Fig. 7). Efficient strong-stop reverse transcription (the first DNA synthesis step) was observed upon incubation of cores with dNTPs. Moreover, addition of DNase I, RNase A or the promiscuous nuclease benzonase failed to prevent encapsidated reverse transcription (Fig. 4a, Extended Data Fig. 7e). This demonstrates that dNTPs can access the interior of the capsid but larger nucleases cannot, supporting the notion of a size-selective pore. Processivity beyond strong-stop was observed but at lower efficiency, in agreement with published data[22].

If the R18 pore is responsible for dNTP import, it is conceivable that capsid mutations that affect the movement of the β-hairpin may also affect the efficiency of reverse transcription. Of the residues primarily responsible for the hairpin movement (Fig. 1a), P1 and D51 are

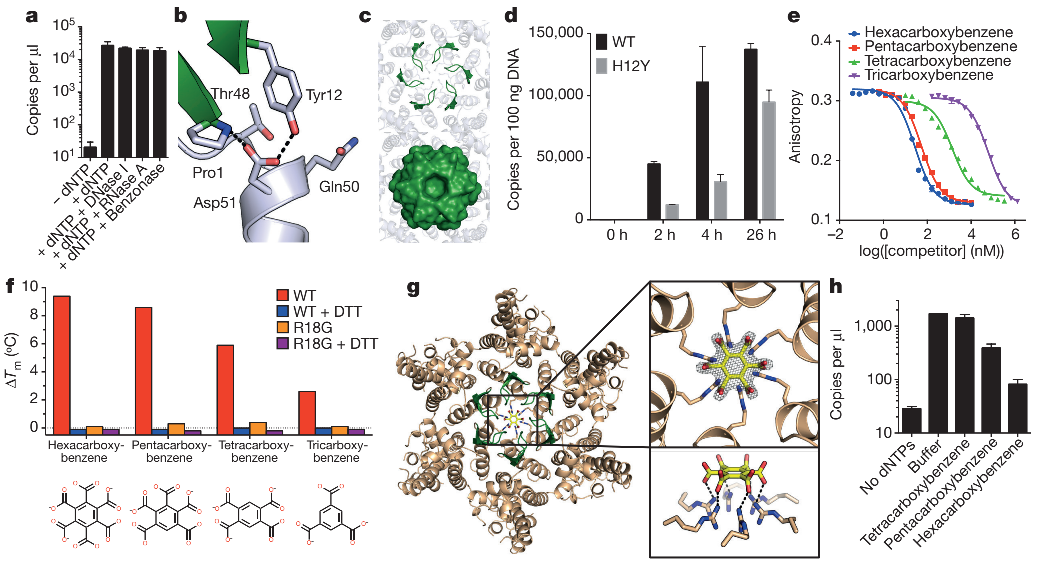


Figure 4 | HIV-1 reverse transcription is inhibited by blockade of the capsid pore. a, *In vitro* endogenous reverse transcription measuring strong-stop transcripts. **b**, Residues surrounding Y12 in the H12Y hexamer structure. **c**, Cartoon and surface representations of the β-hairpin in the H12Y hexamer. **d**, Wild-type and H12Y reverse transcription kinetics. **e**, Competition binding of carboxybenzene compounds to $CA_{hexamer}$. **f**, Change in wild-type and R18G $CA_{hexamer}$ T_m, as measured by DSF in the presence of carboxybenzene compounds. **g**, $CA_{hexamer}$ crystal structure in complex with hexacarboxybenzene, which is co-ordinated by R18. **h**, Effect of carboxybenzene compounds on endogenous reverse transcription. All measurments are triplicate ± s.e.m.

invariant, with mutations at these positions resulting in non-infectious particles due to defective capsid assembly[12]. His12 is also highly conserved, however, in approximately 2% of sequences it has been replaced by a tyrosine. As tyrosine is not titratable over physiological pH values, we hypothesized that the H12Y mutation would result in the β-hairpin favouring one conformation. Solving the crystal structure of this mutant revealed that under the high-pH condition, Y12 displaces the bound water molecule and makes a hydrogen-bond contact with Asp51 (Fig. 4b). Despite contacting D51 directly, the larger side chain of Y12 relative to H12 causes the β-hairpin to favour the closed conformation (Fig. 4c). Notably, H12Y does not completely shut the pore because residues 4–9 do not occupy a single defined state (Extended Data Fig. 8). The β-hairpin therefore retains a degree of flexibility despite rigidification about the P1–D51 'hinge'. Nevertheless, we observed that favouring the closed conformation resulted in H12Y having reduced reverse transcription kinetics while retaining some infectivity (Fig. 4d).

To provide further evidence that nucleotides are recruited through the R18 pore to allow ERT, we sought a small molecule inhibitor that would block the pore. Small polyanionic compounds have been used previously to block analogous arginine-rich pores[23]. We found that the hexacarboxybenzene series (which are polyanionic at physiological pH) bound to $CA_{hexamer}$ and competed for nucleotide binding, as measured by fluorescence anisotropy and DSF, respectively (Fig. 4e, f, Extended Data Fig. 3). Activity broadly increased with the number of negative charges present within the compound, with hexa- or pentacarboxybenzene being the most effective. DSF indicated that the compounds did not bind in the absence of R18 or when hexamers were reduced by DTT, and the crystal structure of the hexacarboxybenzene-bound $CA_{hexamer}$ confirmed that the compound was co-ordinated by R18 within the central pore (Fig. 4f, g). At sufficiently high concentration, tetra-, penta- and hexacarboxybenzene fully inhibit reverse transcriptase, presumably by competing with dNTPs (Extended Data Fig. 7f). However, in ERT assays, when reverse transcriptase is enclosed within an intact viral capsid, tetracarboxybenzene has no effect on reverse transcription, with only a small effect observed for pentacarboxybenzene (Fig. 4h). In contrast, hexacarboxybenzene inhibited ERT almost completely. The failure of tetracarboxybenzene to inhibit ERT demonstrates that a compound sufficiently small to pass through the channel is still efficiently excluded from the capsid interior if it cannot bind the pore. This result emphasizes the chemical selectivity of the pore and its role in dNTP import during reverse transcription.

As a semi-permeable reaction chamber, the HIV-1 capsid is reminiscent of bacterial microcompartments—primitive 'organelles' that utilize a protein coat to isolate toxic reaction intermediates from the cytoplasm[24]. Microcompartments import substrates through a size-selective pore to be consumed by enzymes located inside a chamber[25]. Similarly, dNTPs translocated inside the HIV capsid will be hydrolysed by encapsidated reverse transcriptase. Coupling import with hydrolysis may create a local chemical gradient, promoting interior movement despite the release of captured dNTPs on either side of the pore. Appositely, a subset of microcompartments, carboxysomes, contain positively charged amino acids that are thought to selectively transport bicarbonate and ribulose-1,5-bisphosphate over uncharged CO_2 and O_2[26]. Some microcompartment structures have gated channels located at the six-fold axes in their protein lattice, to control substrate entry and product release. The fact that a similar 'gate' potentially exists in the HIV-1 capsid, provided by the 'molecular iris' of the β-hairpin, suggests that the virus could use this as a mechanism to regulate reverse transcription. In addition, the regulation of capsid stability through dNTP recruitment, and possibly DNA synthesis, provides a model whereby HIV-1 may co-regulate DNA synthesis and uncoating to facilitate cytoplasmic DNA synthesis that remains invisible to cytoplasmic DNA sensing. Finally, the high degree of conservation of R18 coupled with the fact that the pore can be obstructed chemically identifies the pore as a novel target for drug development.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 8 January; accepted 12 July 2016.
Published online 10 August 2016.

1. Campbell, E. M. & Hope, T. J. HIV-1 capsid: the multifaceted key player in HIV-1 infection. *Nat. Rev. Microbiol.* **13,** 471–483 (2015).
2. Rasaiyaah, J. *et al.* HIV-1 evades innate immune recognition through specific cofactor recruitment. *Nature* **503,** 402–405 (2013).
3. Price, A. J. *et al.* Host cofactors and pharmacologic ligands share an essential interface in HIV-1 capsid that is lost upon disassembly. *PLoS Pathog.* **10,** e1004459 (2014).

4. Arhel, N. J. *et al.* HIV-1 DNA Flap formation promotes uncoating of the pre-integration complex at the nuclear pore. *EMBO J.* **26,** 3025–3037 (2007).
5. Gamble, T. R. *et al.* Crystal structure of human cyclophilin A bound to the amino-terminal domain of HIV-1 capsid. *Cell* **87,** 1285–1294 (1996).
6. Kelly, B. N. *et al.* Implications for viral capsid assembly from crystal structures of HIV-1 Gag(1-278) and CA(N)(133-278). *Biochemistry* **45,** 11257–11266 (2006).
7. Ylinen, L. M. J. *et al.* Conformational adaptation of Asian macaque TRIMCyp directs lineage specific antiviral activity. *PLoS Pathog.* **6,** e1001062 (2010).
8. Pornillos, O. *et al.* X-ray structures of the hexameric building block of the HIV capsid. *Cell* **137,** 1282–1292 (2009).
9. Du, S. *et al.* Structure of the HIV-1 full-length capsid protein in a conformationally trapped unassembled state induced by small-molecule binding. *J. Mol. Biol.* **406,** 371–386 (2011).
10. Pornillos, O., Ganser-Pornillos, B. K. & Yeager, M. Atomic-level modelling of the HIV capsid. *Nature* **469,** 424–427 (2011).
11. Price, A. J. *et al.* CPSF6 defines a conserved capsid interface that modulates HIV-1 replication. *PLoS Pathog.* **8,** e1002896 (2012).
12. von Schwedler, U. K. *et al.* Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO* **J.17,** 1555–1568 (1998).
13. Kennedy, E. M., Amie, S. M., Bambara, R. A. & Kim, B. Frequent incorporation of ribonucleotides during HIV-1 reverse transcription and their attenuated repair in macrophages. *J. Biol. Chem.* **287,** 14280–14288 (2012).
14. Bar-Even, A., Milo, R., Noor, E. & Tawfik, D. S. The Moderately Efficient Enzyme: Futile Encounters and Enzyme Floppiness. *Biochemistry* **54,** 4969–4977 (2015).
15. Schreiber, G. & Fersht, A. R. Rapid, electrostatically assisted association of proteins. *Nat. Struct. Biol.* **3,** 427–431 (1996).
16. Magalhaes, A., Maigret, B., Hoflack, J., Gomes, J. N. & Scheraga, H. A. Contribution of unusual arginine-arginine short-range interactions to stabilization and recognition in proteins. *J. Protein Chem.* **13,** 195–215 (1994).
17. Neves, M. A., Yeager, M. & Abagyan, R. Unusual arginine formations in protein function and assembly: rings, strings, and stacks. *J. Phys. Chem. B* **116,** 7006–7013 (2012).
18. Mortuza, G. B. *et al.* Structure of B-MLV capsid amino-terminal domain reveals key features of viral tropism, gag assembly and core formation. *J. Mol. Biol.* **376,** 1493–1508 (2008).
19. Rihn, S. J. *et al.* Extreme genetic fragility of the HIV-1 capsid. *PLoS Pathog.* **9,** e1003461 (2013).
20. Keckesova, Z., Ylinen, L. M. & Towers, G. J. The human and African green monkey TRIM5alpha genes encode Ref1 and Lv1 retroviral restriction factor activities. *Proc. Natl Acad. Sci. USA* **101,** 10780–10785 (2004).
21. Shah, V. B. & Aiken, C. *In vitro* uncoating of HIV-1 cores. *J. Vis. Exp.* **57,** 3384 (2011).
22. Warrilow, D., Warren, K. & Harrich, D. Strand transfer and elongation of HIV-1 reverse transcription is facilitated by cell factors *in vitro*. *PLoS One* **5,** e13229 (2010).
23. Cheley, S., Gu, L. Q. & Bayley, H. Stochastic sensing of nanomolar inositol 1,4,5-trisphosphate with an engineered pore. *Chem. Biol.* **9,** 829–838 (2002).
24. Tanaka, S., Sawaya, M. R. & Yeates, T. O. Structure and mechanisms of a protein-based organelle in *Escherichia coli*. *Science* **327,** 81–84 (2010).
25. Chowdhury, C. *et al.* Selective molecular transport through the protein shell of a bacterial microcompartment organelle. *Proc. Natl Acad. Sci. USA* **112,** 2990–2995 (2015).
26. Kerfeld, C. A. *et al.* Protein structures forming the shell of primitive bacterial organelles. *Science* **309,** 936–938 (2005).

Supplementary Information is available in the online version of the paper.

Acknowledgements This work was funded by the Medical Research Council (UK; U105181010), the European Research Council (281627 -IAI), the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013), ERC grant agreement number 339223 and the National Institute for Health Research University College London Hospitals Biomedical Research Centre. G.J.T. was supported by a Wellcome Trust Senior Biomedical Research Fellowship, D.A.J. by an NHMRC Early Career Fellowship (CJ Martin) (GNT1036521) and AJP by a Research Fellowship from Emmanuel College, Cambridge. We thank L. McKeane for her help designing figures.

Author Contributions D.A.J. performed the majority of the protein production, crystallization experiments and analysis; the fluorescence anisotropy binding experiments; differential scanning fluorimetry; chimaeric virus production and associated infectivity and RT measurements; and TRIM5 abrogation assay. W.A.M. performed the HIV core preparation and endogenous RT experiments. L.H. performed R18G and H12Y infectivity and RT characterizations. A.J.P. crystallized and collected diffraction data from CA$_{\text{hexamer}}$ in the open state. L.C.J. performed the stopped-flow kinetics experiments. G.J.T. and L.C.J. supervised the project. The paper was primarily written by D.A.J. and L.C.J. All authors discussed the results and implications and commented on the manuscript at all stages.

Author Information Atomic coordinates and structure factor files have been deposited in the Protein Data Bank under accession numbers 5HGK, 5HGL, 5HGM, 5HGN, 5HGO, and 5JPA. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to LCJ (lcj@mrc-lmb.cam.ac.uk) or GJT (g.towers@ucl.ac.uk).

Reviewer Information *Nature* thanks H. Bayley, P. Cherepanov, M. Yeager and T. Yeates for their contribution to the peer review of this work.

METHODS

Protein production and purification. The CA N-terminal domain and the disulfide-stabilized $CA_{hexamer}$ were expressed and purified as previously described[8,27]. The R18G mutation was introduced by QuikChange site-directed mutagenesis. Chimaeric $CA_{hexamers}$ were produced by mixing the desired ratio of pre-assembled wild-type and R18G $CA_{hexamer}$ ($16\,mg\,ml^{-1}$) followed by a four-step dialysis: (1) disassembly in Tris (pH 8.0, 50 mM), NaCl (40 mM), β-mercaptoethanol (20 mM); (2) reassembly in Tris (pH 8.0, 50 mM), NaCl (1 M), β-mercaptoethanol (20 mM); (3) oxidation in Tris (pH 8.0, 50 mM), NaCl (1 M); (4) redispersion in Tris (pH 8.0, 20 mM), NaCl (40 mM). In the context of the chimaera experiments, wild type and R18G were also subjected to this process so that samples were matched with the other ratios. Reassembled hexamers were observed by non-reducing SDS–PAGE. Chimaeric hexamers were compared with mixes of homohexamers by fluorescence anisotropy (see below and Extended Data Fig. 5) in order to demonstrate that chimaeras had indeed formed.

Crystallization, structure solution and analysis. All crystals were grown at 17 °C by sitting-drop vapour diffusion in which 100 nl protein was mixed with 100 nl precipitant and suspended above 80 μl precipitant. The CA N-terminal domain ($15\,mg\,ml^{-1}$) 'open' conformation was crystallized from PEG3350 (20%), ammonium chloride (0.2 M, pH 6.3). Crystals were cryoprotected in precipitant supplemented with 25% glycerol. The $CA_{hexamer}$ ($15\,mg\,ml^{-1}$) 'open conformation' was crystallized from PEG4000 (12%), NaCl (0.1 M), $MgCl_2$ (0.1 M), sodium citrate (0.1 M, pH 5.5). Crystals were cryoprotected in precipitant supplemented with 20% MPD. The remaining $CA_{hexamer}$ structures (apo, dATP-bound, R18G and hexacarboxybenzene-bound) were all obtained from 10–12 $mg\,ml^{-1}$ protein mixed with PEG550MME (13–14%), KSCN (0.15 M), Tris (0.1 M, pH 8.5) and cryoprotected with precipitant supplemented with 20% MPD. For the dATP-bound structure, the protein was supplemented with 10 mM dATP immediately before crystallization; whereas for the hexacarboxybenzene structure, the protein was likewise supplemented with 1 mM hexacarboxybenzene (Tris-buffered to pH 8.0). All crystals were flash-cooled in liquid nitrogen and data collected either in-house using Cu $K\alpha$ X-rays produced by a Rigaku FR-E rotating anode generator with diffraction recorded on a mar345 image plate detector (marXperts), or at beamline I02 at Diamond Light Source. The data sets were processed using the CCP4 program suite[28]. Data were indexed and integrated with IMOSFLM[29] and scaled and merged with either POINTLESS and SCALA[30] or AIMLESS[31]. Structures were solved by molecular replacement using PHASER[32] and refined using REFMAC5[33]. Between rounds of refinement, the model was manually checked and corrected against the corresponding electron-density maps in COOT[34]. Solvent molecules and bound ligands were added as the refinement progressed either manually or automatically within COOT, and were routinely checked for correct stereochemistry, for sufficient supporting density above a $2F_o - F_c$ threshold of 1.0σ and for a reasonable thermal factor. The quality of the model was regularly checked for steric clashes, incorrect stereochemistry and rotamer outliers using MOLPROBITY[35]. Final figures were rendered in The PyMOL Molecular Graphics System, Version 1.5.0.4 Schrödinger, LLC. Surface electrostatics were calculated using the APBS PyMOL plugin[36] and cavity volume measurements with 3V[37]. Data collection and refinement statistics are presented in Extended Data Table 2. The 'fullerene cone' model of an HIV-1 virion is based on 3J3Q[38], but using the inter-hexamer packing from 4XFY[39] and an open β-hairpin conformation.

Fluorescence anisotropy. Fluorescence anisotropy measurements were performed at 22 °C on a Cary Eclipse Fluorescence Spectrophotometer (Agilent). Fluorescein-labelled dNTPs were obtained from Perkin Elmer and used for saturation binding experiments at a concentration of 2 nM prepared in 'Intracellular Buffer': potassium gluconate (110 mM), KCl (25 mM), NaCl (5 mM), $MgCl_2$ (2 mM), HEPES (10 mM), final pH 7.2. $CA_{hexamer}$ disassembly was achieved by the addition of DTT (4 mM), and was performed routinely at the conclusion of each saturation binding experiment to confirm the absence of non-specific binding. It was found that the triphosphate was not stable over the timescale of the competition binding experiments; so fluorescein-labelled dNTPs were substituted for a non-hydrolysable BODIPY-labelled GTP-γ-S (ThermoFisher Scientific). Saturation binding experiments determined that this non-hydrolysable analogue bound with unchanged affinity to the $CA_{hexamer}$ ($K_d = 14$ nM). 200 mM stock solutions of hexacarboxybenzene (Sigma), pentacarboxybenzene (MP Biomedicals), 1,2,4,5-tetracarboxybezene (Sigma), and 1,3,5-tricarboxybenzene (Fluka) were prepared in 50 mM Tris and adjusted to pH 8.0. For competition binding experiments, the competitor was titrated into a mix of $CA_{hexamer}$ (28 nM) and BODIPY-GTP-γ-S (2 nM). All fluorescence anisotropy measurements are representative of at least two experiments. Each point is measured in quadruplicate and plotted as mean ± standard deviation. In many cases, error bars lie within the datapoint. Saturation binding and competition binding curves were fit using GraphPad Prism (GraphPad Software, Inc.).

Rapid reaction kinetics. Experiments were carried out using a dual-channel fluorescence TgK single-mix SF-61SX2 stopped-flow spectrometer. All samples were prepared in Intracellular Buffer. Mixing was performed 1:1, using an excitation wavelength of 488 nm and a 520 nm cut-off filter. Association experiments were carried out at 0.25 μM dCTP and a range of μM $CA_{hexamer}$ concentrations. Dissociation experiments were carried out using 20 μM unlabelled dCTP and a pre-formed fluorescein-labelled 1 μM dCTP:$CA_{hexamer}$ complex. Relaxation rates were determined using a single exponential model: $F = \Delta F_{exp}(-k_{obs}t) + F_e$, where F is the observed fluorescence, ΔF is the fluorescence amplitude, k_{obs} is the observed pseudo first-order rate constant, and F_e is the end-point fluorescence. The bimolecular association rate constant (k_{on}) was determined by fitting the linear relationship between k_{obs} and the increasing pseudo-first order concentrations of $CA_{hexamer}$ to: $k_{obs} = k_{on}[CA_{hexamer}] + k_{reverse}$. For stopped flow experiments, every 0.5 s measurement included >2,000 datapoints, each of which was oversampled 99 times. At least three independent mixing experiments were averaged for each ligand concentration.

Differential scanning fluorimetry. DSF measurements were performed using a Prometheus NT.48 (NanoTemper Technologies) over a temperature range of 20–95 °C using a ramp rate of $2.5\,°C\,min^{-1}$. $CA_{hexamer}$ samples were prepared at a final concentration of $1\,mg\,ml^{-1}$ in Intracellular Buffer (±DTT (4 mM)). dNTPs or competitors were added at 200 μM. DSF scans are single reads. Consistency between like points yields an uncertainty in T_m of no greater than 0.2 °C.

Cells and viruses. All cell lines were obtained from ATCC and tested negative for mycoplasma contamination. Replication deficient VSV-G pseudotyped HIV GFP vectors were produced in HEK293T cells as described previously[3]. Site-directed mutagenesis of CA was performed using the QuikChange method (Stratagene) against the Gag-Pol expression plasmid, pCRV-1. Chimaeric viruses were produced by mixing the appropriate ratio of wild-type or mutant pCRV-1 before transfection. Reverse transcriptase activity was quantified using a colorimetric ELISA assay (Roche) and was found not to vary significantly between viruses. Production of mature particles was confirmed by western blot for p24 from pelleted virus, with no observable difference between chimaeras. Primary antibody used for western blotting was a polyclonal goat anti-p24 (Bio-Rad, product 4999-9007).

Infection experiments. Infections of HeLa cells were performed in the presence of $5\,\mu g\,ml^{-1}$ polybrene. GFP expressing cells were enumerated on a BD LSRII flow cytometer (BD Biosciences) 2 days post-transfection after fixation of cells in 4% paraformaldehyde. Chimaera infectivity was determined by a six-point titration of each chimaera onto HeLa cells. Values are the mean ± standard deviation calculated from all points for which the proportion of infected cells after 48 h was between 1% and 50%.

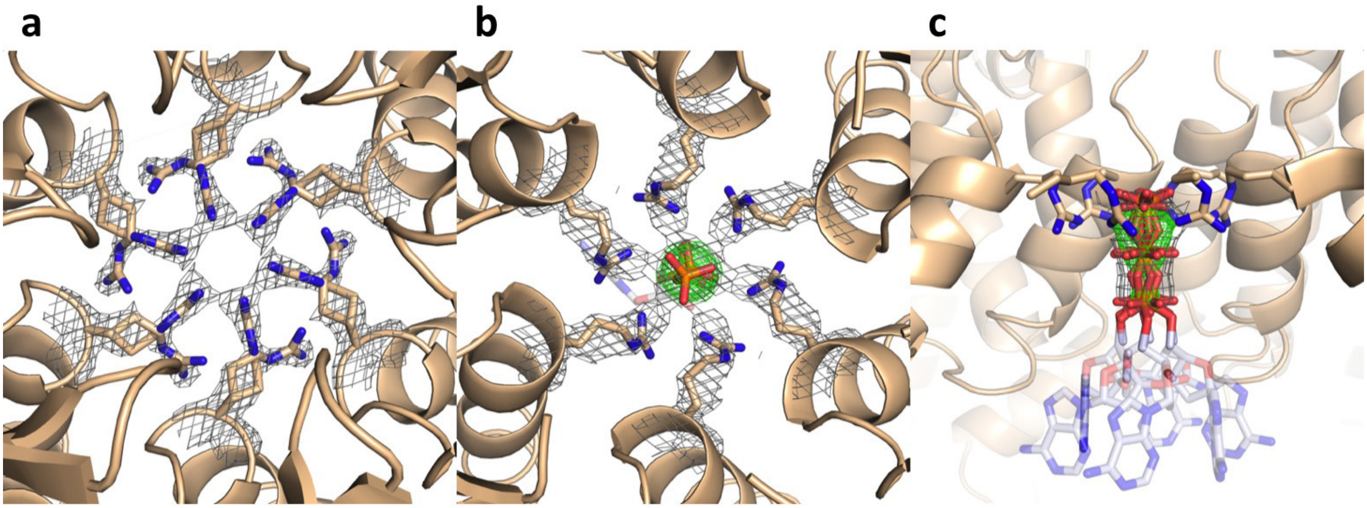
TRIM5α abrogation assay. 'Abrogating virus' (VSV-G pseudotyped HIV puromycin vectors) was produced as described above, with the exception of the *gfp* gene, which was replaced with the *pac* gene to ensure that the virus did not confer fluorescence upon infection. Virus was concentrated by ultracentrifugation with an SW28 rotor at 25,000 rpm (112,700*g*) for 2 h. The abrogating virus capsids were wild type, R18G or W184A/M185A (a mutant with a known assembly defect that cannot compete for TRIM5α). VSV-G pseudotyped HIV GFP vectors were titrated on FRhK-4 cells in the presence of $5\,\mu g\,ml^{-1}$ polybrene to determine the volume of virus required to achieve 1% infection. In a separate experiment, cells were then co-infected with that amount HIV–GFP vector and a titration of VSV-G pseudotyped HIV puromycin vectors (the abrogating virus). GFP-expressing cells were measured in duplicate and enumerated as above. Results are representative of three experiments and are presented as mean ± standard deviation. For many points the error bars lie within the datapoint.

Quantitative PCR. For analysis of reverse transcription products, viral supernatant was treated with $250\,U\,ml^{-1}$ DNase (Millipore) for 1 h before infection. Cells were collected 6 h after infection. DNA was extracted using DNeasy Blood and Tissue Kit (Qiagen). GFP copies were quantified using primers GFPF (5'-CAACAGCCACAACGTCTATATCAT-3'), GFPR (5'-ATGTTGTGGCGGATCTTGAAG-3') and probe GFPP (5'-(FAM) CCGACAAGCAGAAGAACGGCATCAA(TAMRA)-3') against a standard curve of CSGW on an ABI StepOnePlus Real Time PCR System (Life Technologies). Chimaera reverse transcription measurements are representative of three experiments with each point measured in triplicate. Results are presented as mean ± standard deviation. For H12Y, a time course was also performed, in which each time point was measured in triplicate and presented as above.

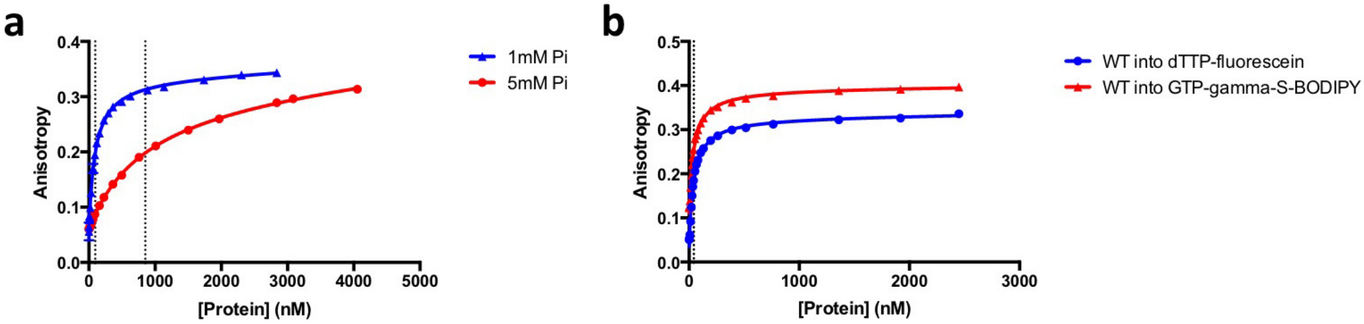
Preparation of HIV-1 cores. HIV-1 capsid cores were prepared using a protocol based on ref. 21 with modifications. 90 ml HEK293T supernatant containing VSV-G pseudotyped HIV-1 GFP was pelleted over 20% sucrose dissolved in core prep buffer (CPB; 20 mM Tris (pH 7.4), 20 mM NaCl, 1 mM $MgCl_2$) in an SW28 rotor (Beckman) at 25,000 rpm at 4 °C. Pellets were gently resuspended at 4 °C in CPB for 1 h with occasional agitation. Resuspended pellets were treated with DNase I from bovine pancreas (Sigma Aldrich) for 1 h at $200\,\mu g\,ml^{-1}$ at room temperature to remove contaminating extra-viral DNA. Virus was subjected to spin-through detergent stripping of the viral membrane as follows. A gradient at 80–30% sucrose was prepared in SW40Ti ultracentrifuge tubes and overlaid with 250 μl 1% Triton X-100 in 15% sucrose, followed by 250 μl 7.5% sucrose. All solutions were prepared in CPB. 750 μl DNase-treated, concentrated virus was layered on top of the gradient and subjected to 32,500 rpm at 4 °C for 16 h. The preparation was fractionated and the location of cores was determined by ELISA for p24 (Perkin Elmer). Core-containing fractions were pooled and snap frozen before storage at −80 °C.

Endogenous reverse transcription assays. Viral cores were diluted to 400 μg ml^{-1} p24 with 60% sucrose in CPB and pre-treated with nucleases for 1 h before addition of dNTPs. Final concentrations of dNTPs were 100 μM each, DNase I and RNase A were at 100 μg ml^{-1} and benzonase was at 250 U ml^{-1}. 20 μl reactions were incubated at room temperature for 16 h unless indicated otherwise and were stopped by shifting to −80 °C. DNA was prepared using DNeasy Blood and Tissue kit (Qiagen) after addition of 200 μl PBS with of 50 μg ml^{-1} salmon sperm carrier DNA to each sample. Reverse transcript products were detected using TaqMan Fast Universal PCR Mix (ABI) and RU5 primers to detect strong-stop DNA[40] (RU5 forward: 5'-TCTGGCTAACTAGGGAACCCA-3'; RU5 reverse: 5'-CTGACTAAAAGGGTCTGAGG-3'; and RU5 probe 5'-(FAM) TTAAGCCTCAATAAAGCTTGCCTTGAGTGC(TAMRA)-3'), GFP primers to detect first-strand transfer products (described above) and primers for second-strand transfer products[40] (2ST forward: 5'-TTTTAGTCAGTGTGGAAAATCTGTAGC-3'; 2ST reverse: 5'-TACTCACCAGTCGCCGCC-3'; and 2ST probe: 5'-(FAM) TCGACGCAGGACTCGGCTTGCT(TAMRA)-3'). Where used, carboxybenzene compounds were dissolved in CPB, pH-adjusted with NaOH and added to reactions at a final concentration of 20 mM. For dNTP concentration to be limiting, these reactions were performed in the presence of 1 μM each dNTP and reactions were stopped 5 h after their addition. ERT experiments were performed in experimental triplicate and are representative of several experimental replicates. Data are represented as mean ± s.e.m.

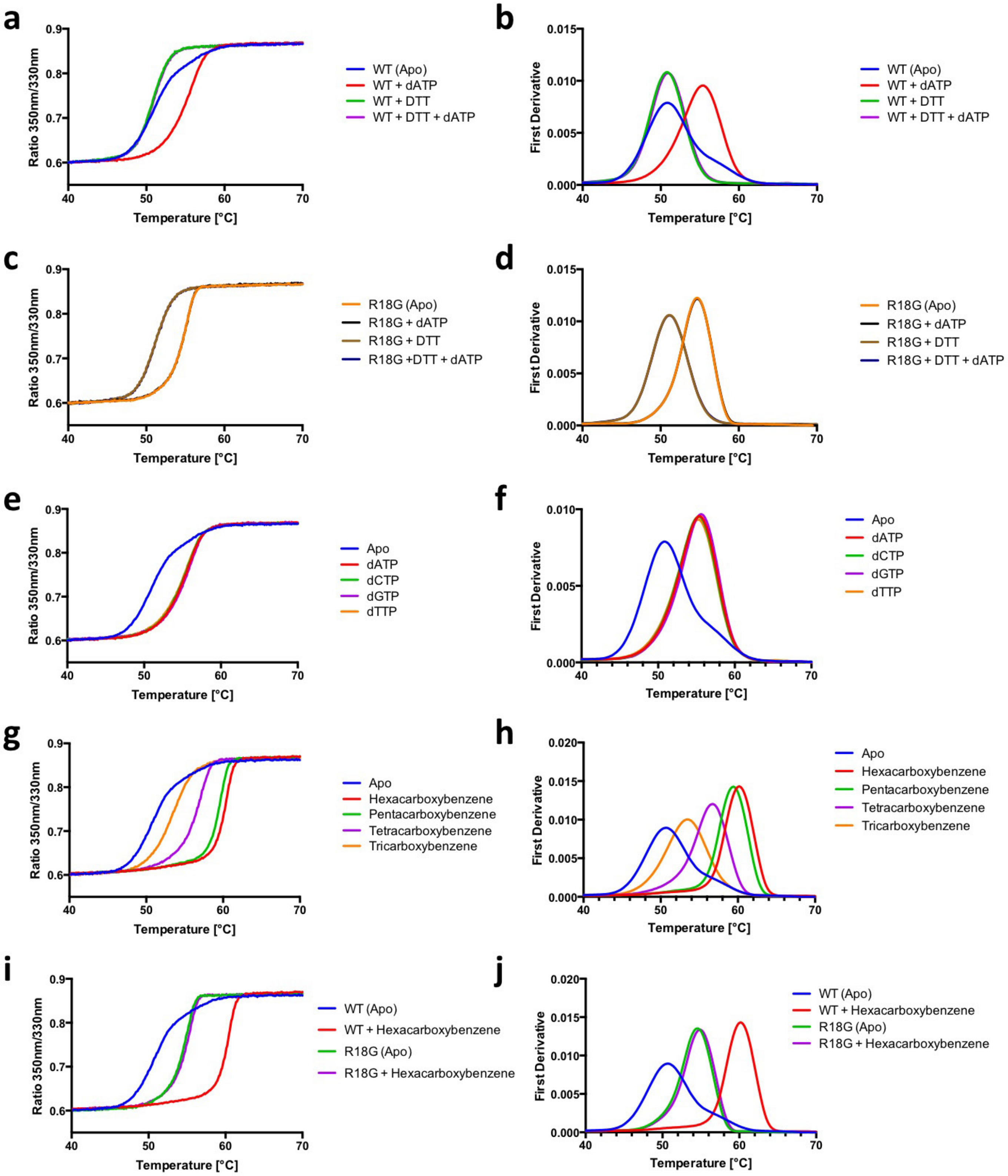
27. Price, A. J. *et al.* Active site remodeling switches HIV specificity of antiretroviral TRIMCyp. *Nat. Struct. Mol. Biol.* **16,** 1036–1042 (2009).
28. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. D* **67,** 235–242 (2011).
29. Leslie, A. G. W. & Powell, H. R. Processing diffraction data with MOSFLM. *Nato Sci Ser Ii Math* **245,** 41–51 (2007).
30. Evans, P. R. An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Crystallogr. D* **67,** 282–292 (2011).
31. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69,** 1204–1214 (2013).
32. McCoy, A. J. *et al.* Phaser crystallographic software. *J. Appl. Crystallogr.* **40,** 658–674 (2007).
33. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53,** 240–255 (1997).
34. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60,** 2126–2132 (2004).
35. Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66,** 12–21 (2010).
36. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA* **98,** 10037–10041 (2001).
37. Voss, N. R. & Gerstein, M. 3V: cavity, channel and cleft volume calculator and extractor. *Nucleic Acids Res.* **38,** W555–W562 (2010).
38. Zhao, G. *et al.* Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497,** 643–646 (2013).
39. Gres, A. T. *et al.* Structural virology. X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability. *Science* **349,** 99–103 (2015).
40. Julias, J. G., Ferris, A. L., Boyer, P. L. & Hughes, S. H. Replication of phenotypically mixed human immunodeficiency virus type 1 virions containing catalytically active and catalytically inactive reverse transcriptase. *J. Virol.* **75,** 6537–6546 (2001).



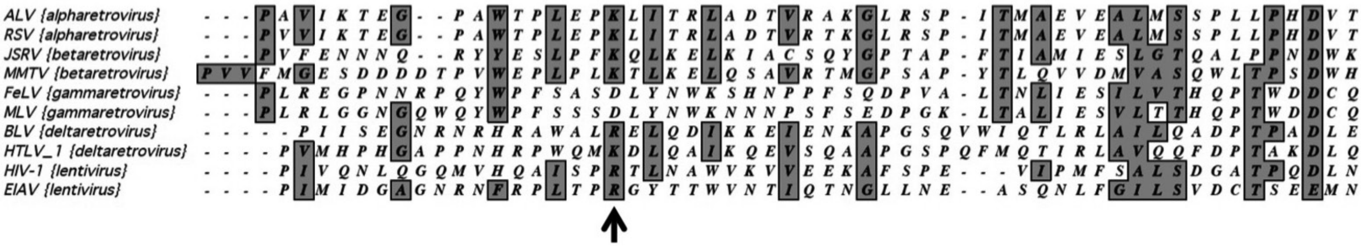
Extended Data Figure 1 | dATP binds to the R18 pore at the centre of the capsid hexamer. a, **b**, $2F_o - F_c$ density (grey mesh) contoured at 1.0σ about R18 for the unbound (**a**) and dATP-bound (**b**) $CA_{hexamer}$ structures. $F_o - F_c$ omit density (green mesh) contoured at 3.0σ is shown for the dATP-bound structure. **c**, dATP lies on the crystallographic six-fold axis and clear rotationally averaged density is observed only for the triphosphate group.



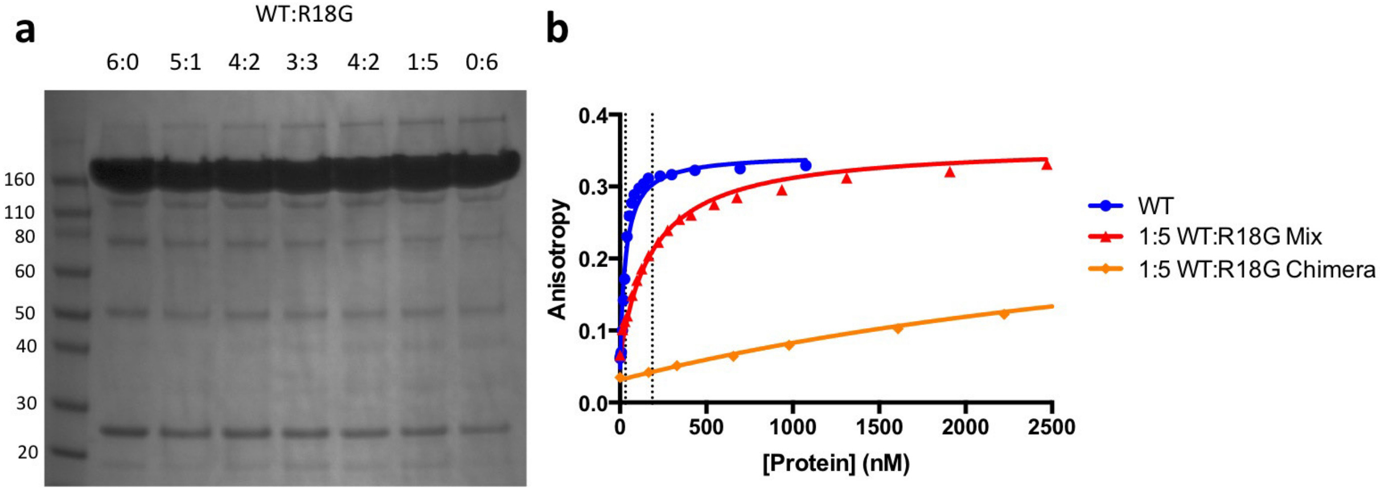
Extended Data Figure 2 | Controls for dNTP-binding experiment. **a**, Titration of $\mathrm{CA_{hexamer}}$ into 2 nM fluorescein-labelled dTTP in the presence of 1 mM (physiological) or 5 mM inorganic phosphate. Under the 1 mM conditions, there is no significant effect on hexamer binding to dTTP. At 5 mM, apparent affinity is decreased to 851 nM, demonstrating that inorganic phosphate can compete for the pore. However, given that the intracellular [dNTP] is approximately 100 μM, under intracellular conditions dNTP binding would dominate. **b**, Titration of $\mathrm{CA_{hexamer}}$ into BODIPY-labelled rGTP-γ-S and fluorescein-labelled dTTP. Each binds with the same affinity, which suggests that the R18 pore is unable to discriminate between ribose and deoxyribose nucleoside triphosphates. The difference in the magnitude of the fluorescence anisotropy signals is due to differences in fluorophore excited state lifetimes. K_D values are indicated by a dotted line. All measurements were performed in quadruplicate and reported as mean ± s.d.



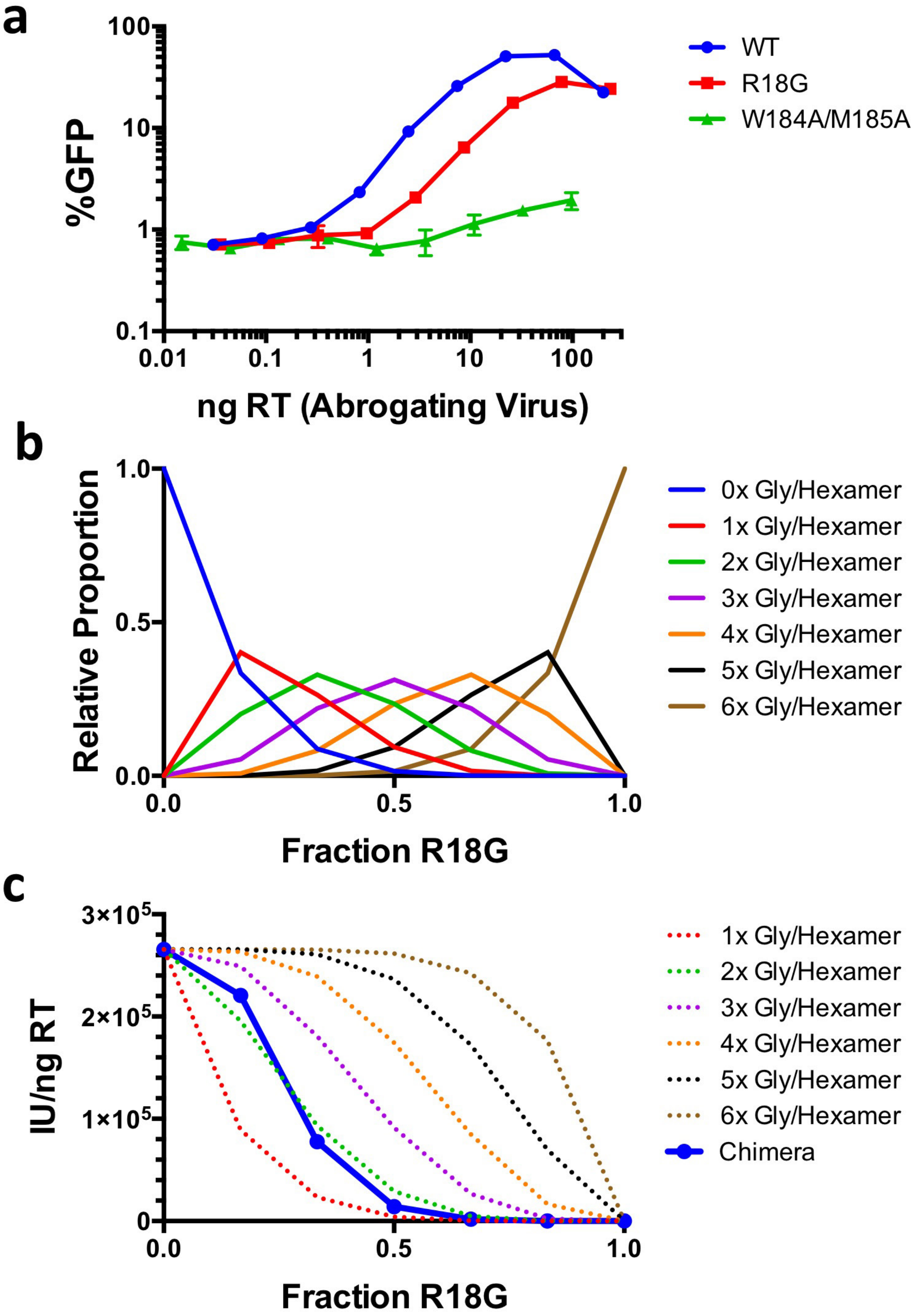
Extended Data Figure 3 | DSF melt curves. The left-hand panels show the ratio of tryptophan fluorescence emission at 350 nm and 330 nm as a function of temperature. The right-hand panels show the first derivative of the same data, the peak of which is used to determine the T_m value. **a**, **b**, Effect of dATP and DTT on wild-type $CA_{hexamer}$. **c**, **d**, Effect of dATP and DTT on R18G $CA_{hexamer}$. **e**, **f**, Effect of each dNTP on wild-type $CA_{hexamer}$. **g**, **h**, Comparison of the effects of carboxybenzene compounds on wild-type $CA_{hexamer}$. **i**, **j**, Comparison of the effects of hexacarboxybenzene on wild-type and R18G $CA_{hexamer}$.


ALV {alpharetrovirus} - - - P A V I K T E G - - P A W T P L E P K L I T R L A D T V R A K G L R S P - I T M A E V E A L M S S P L L P H D V T
RSV {alpharetrovirus} - - - P V V I K T E G - - P A W T P L E P K L I T R L A D T V R T K G L R S P - I T M A E V E A L M S S P L L P H D V T
JSRV {betaretrovirus} - - - P V F E N N N Q - - R Y Y E S L P F K Q L K E L K I A C S Q Y G P T A P - F T I A M I E S L G T Q A L P P N D W K
MMTV {betaretrovirus} P V V F M G E S D D D D T P V W E P L P L K T L K E L Q S A V R T M G P S A P - Y T L Q V V D M V A S Q W L T P S D W H
FeLV {gammaretrovirus} - - - P L R E G P N N R P Q Y W P F S A S D L Y N W K S H N P P F S Q D P V A - L T N L I E S I L V T H Q P T W D D C Q
MLV {gammaretrovirus} - - - P L R L G G N G Q W Q Y W P F S S S D L Y N W K N N N P S F S E D P G K - L T A L I E S V L T T H Q P T W D D C Q
BLV {deltaretrovirus} - - - - - - P I I S E G N R N R H R A W A L R E L Q D I K K E I E N K A P G S Q V W I Q T L R L A I L Q A D P T P A D L E
HTLV_1 {deltaretrovirus} - - - - - P V M H P H G A P P N H R P W Q M K D L Q A I K Q E V S Q A A P G S P Q F M Q T I R L A V Q Q F D P T A K D L Q
HIV-1 {lentivirus} - - - - - P I V Q N L Q G Q M V H Q A I S P R T L N A W V K V V E E K A F S P E - - - V I P M F S A L S D G A T P Q D L N
EIAV {lentivirus} - - - - - P I M I D G A G N R N F R P L T P R G Y T T W V N T I Q T N G L L N E - - A S Q N L F G I L S V D C T S E E M N

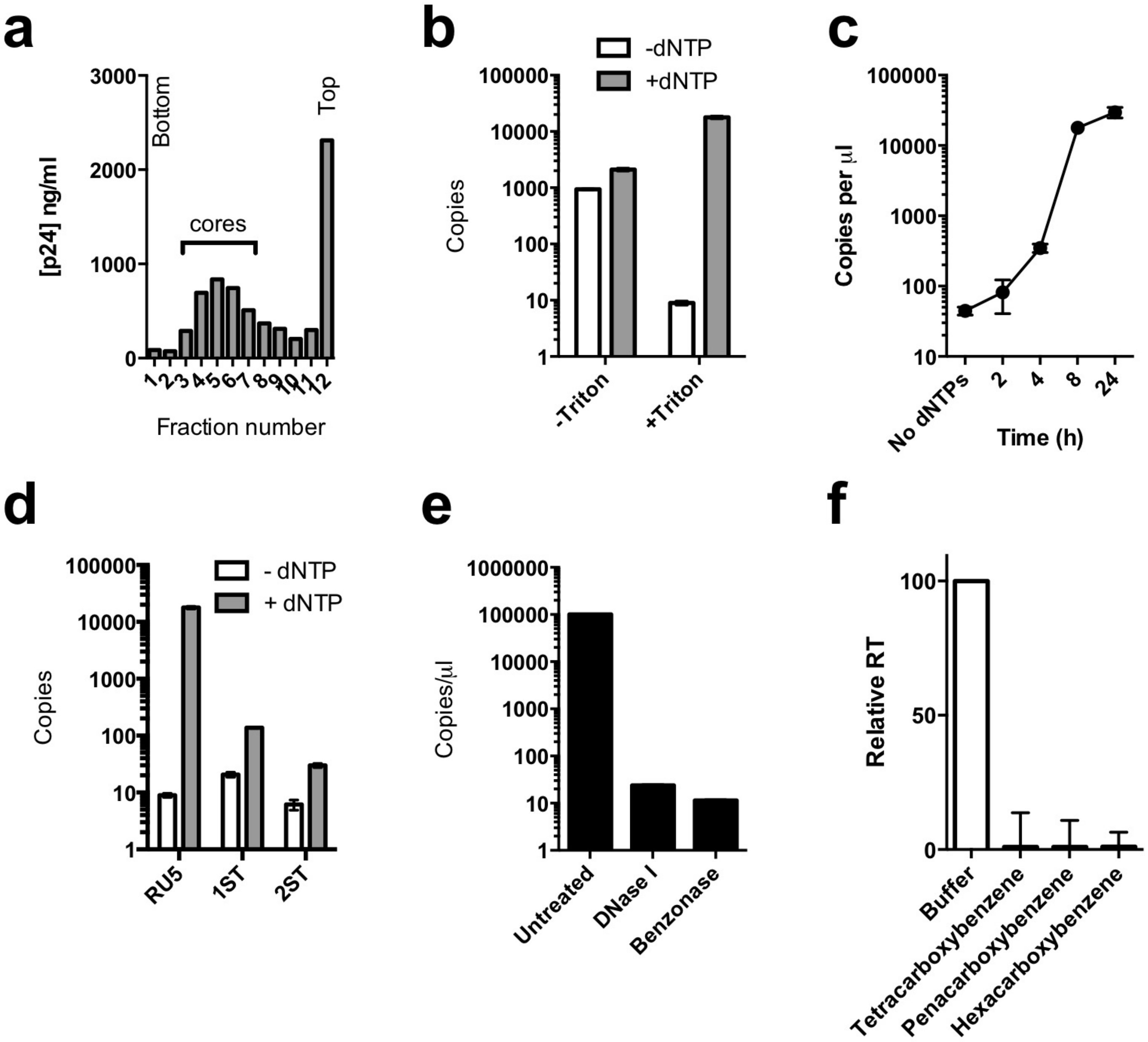
Extended Data Figure 4 | Alignment of selected retrovirus capsid sequences bordering the electropositive pore. The position equivalent to R18 in HIV-1 is marked with an arrow.



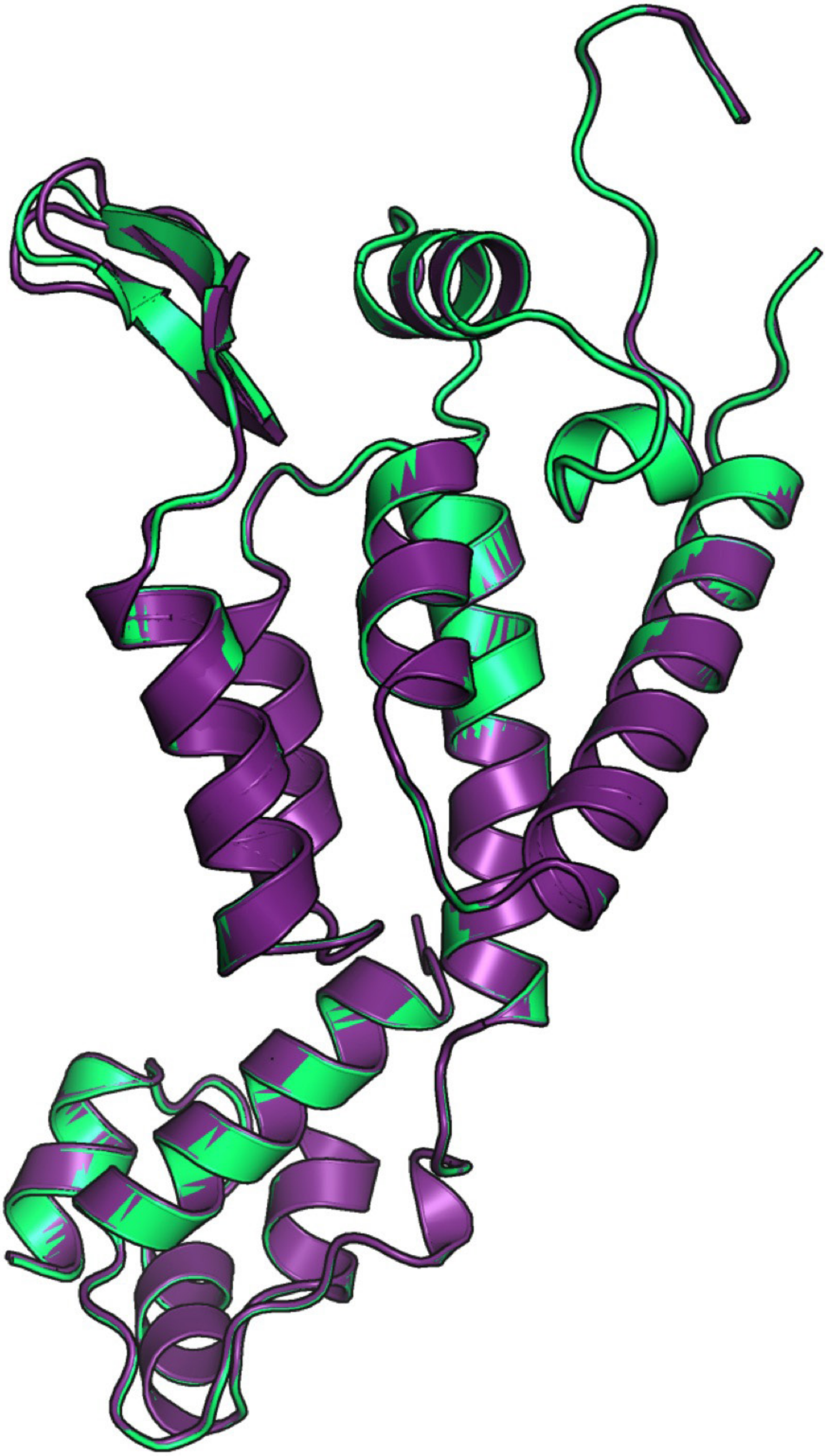
Extended Data Figure 5 | Confirmation of CA$_{\text{hexamer}}$ chimaera assemblies. a, Non-reducing SDS–PAGE of CA$_{\text{hexamer}}$ wild-type:R18G chimaera samples demonstrates that the recombinant proteins had reassembled into hexamers. Molecular weight standards (kDa) are presented in the first lane. For gel source data, see Supplementary Fig. 1. **b**, Comparison of 1:5 homohexamer mix and the equivalent chimaera. There is a six-fold loss of apparent Kd for the wild-type:R18G mix, as expected for a six-fold dilution of wild type with a non-binding mutant. In contrast, the 1:5 chimaera chimaera has a 58-fold decrease in K_d, demonstrating that chimaeric hexamers had indeed formed. All measurements were performed in quadruplicate and reported as mean ± s.d.



Extended Data Figure 6 | Effects of HIV-1 CA R18G on viral infectivity. **a**, R18G is capable of abrogating TRIM5α-mediated restriction. Rhesus TRIM5α provides a potent block to infection of HIV in FRhK-4 cells. Titration of a non-GFP-expressing virus can compete for TRIM5α-binding and relieve the restriction of a GFP-expressing virus only if it delivers an assembled capsid into the cytoplasm. R18G abrogates restriction but W184A/M185A, which is incapable of forming assembled capsids due to loss of the CTD–CTD dimerization interface, does not. Reported values are mean of triplicate ± s.d. **b**, Binomial distribution model for the relative proportion of capsid hexamers carrying a discrete number of glycines at position 18 at defined bulk ratios of wild-type:R18G. **c**, Six models (dotted lines) predicting the effect of replacing arginine 18 with glycines. Each model assumes that a different number of glycines is required to render the pore defective. The data from wild-type:R18G chimaeric virus measurements (solid line) are consistent with a model in which four or more arginines (that is, two or fewer glycines, green) are required to maintain a functional pore.

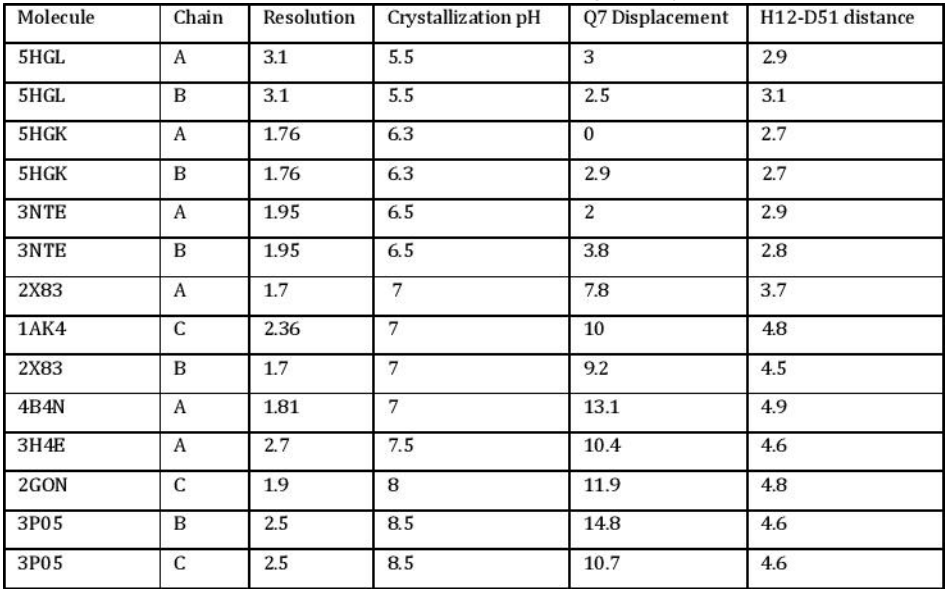


Extended Data Figure 7 | ERT assay. a, HIV-1 cores were prepared by ultracentrifugation through a Triton X-100 layer over a sucrose gradient. Resulting fractions were subjected to ELISA for p24 and fractions 3–7 were pooled for further experiments. **b**, Endogenous reverse transcriptase activity for strong-stop in the presence of DNase I using HIV-1 fractions that were prepared with or without the Triton X-100 spin-through layer. Input levels of p24 were normalized between reactions. **c**, dNTPs were added to HIV-1 cores prepared by Triton X-100 spin-through in the presence of DNase I. Reactions were stopped at the indicated time point by shifting to −80 °C and levels of strong-stop were quantified. **d**, Levels of strong-stop (RU5), first-strand transfer (1ST) and second-strand transfer (2ST) DNA after overnight incubation of HIV-1 cores with or without dNTPs in the presence of DNase I. **e**, Levels of naked HIV-1 DNA genomes untreated or incubated overnight with DNase I or benzonase. **f**, Effect of carboxybenzene compounds on recombinant reverse transcriptase activity. All measurements were performed in triplicate and reported as mean ± s.e.m.



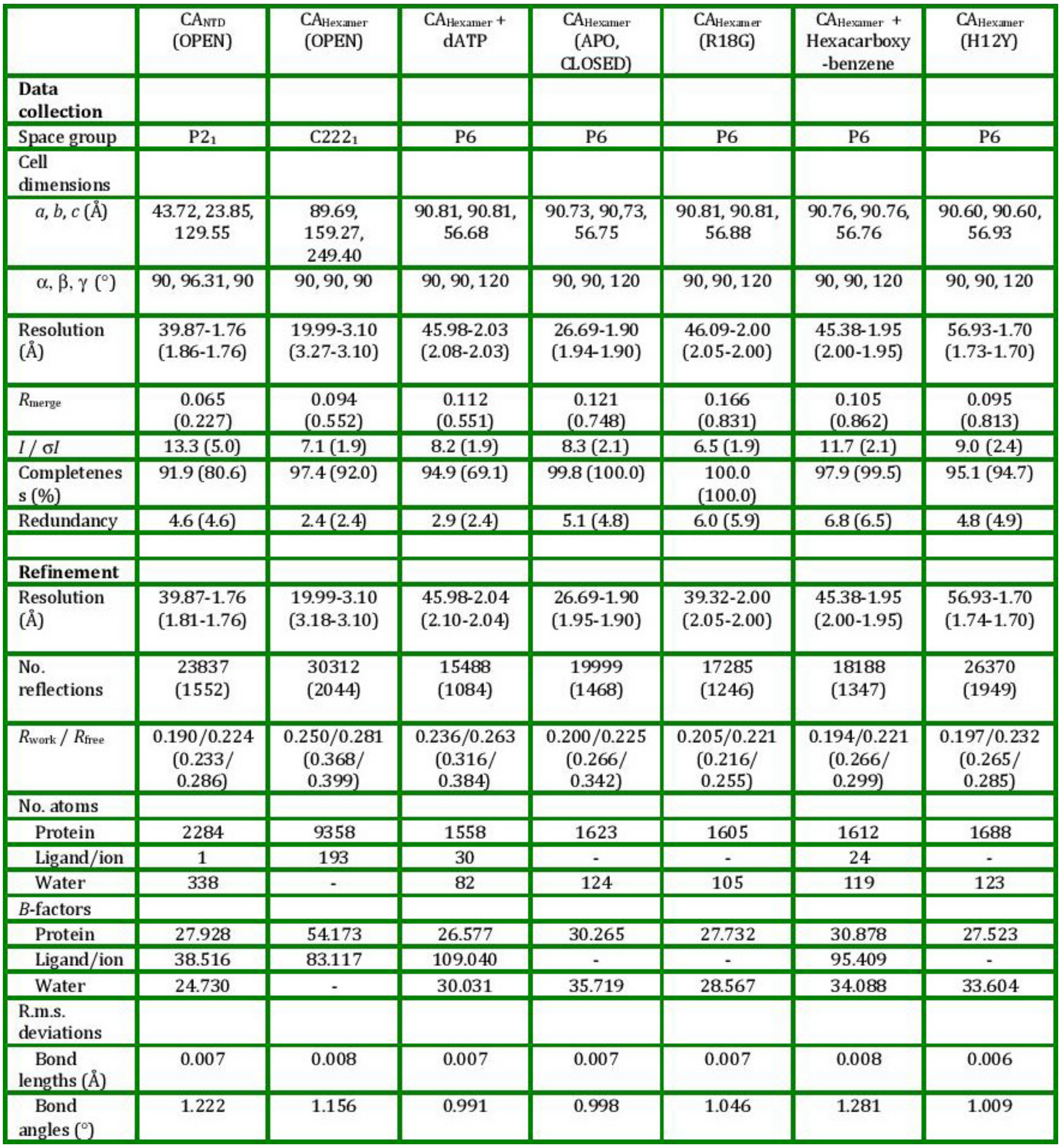
Extended Data Figure 8 | Comparison of wild-type and H12Y crystal structures. The H12Y monomer (in the context of the hexamer, purple) superposes on the wild type (green) with r.m.s.d. of 0.2471 Å. Residues 4–9 of the H12Y structure have been modelled in two alternate conformations, owing to flexibility towards the tip of the hairpin.

Extended Data Table 1 | Details of structures used for β-hairpin analysis



Molecule	Chain	Resolution	Crystallization pH	Q7 Displacement	H12-D51 distance
5HGL	A	3.1	5.5	3	2.9
5HGL	B	3.1	5.5	2.5	3.1
5HGK	A	1.76	6.3	0	2.7
5HGK	B	1.76	6.3	2.9	2.7
3NTE	A	1.95	6.5	2	2.9
3NTE	B	1.95	6.5	3.8	2.8
2X83	A	1.7	7	7.8	3.7
1AK4	C	2.36	7	10	4.8
2X83	B	1.7	7	9.2	4.5
4B4N	A	1.81	7	13.1	4.9
3H4E	A	2.7	7.5	10.4	4.6
2GON	C	1.9	8	11.9	4.8
3P05	B	2.5	8.5	14.8	4.6
3P05	C	2.5	8.5	10.7	4.6

Extended Data Table 2 | Crystallographic data collection and refinement statistics



	CA_{NTD} (OPEN)	$CA_{Hexamer}$ (OPEN)	$CA_{Hexamer}$ + dATP	$CA_{Hexamer}$ (APO, CLOSED)	$CA_{Hexamer}$ (R18G)	$CA_{Hexamer}$ + Hexacarboxy-benzene	$CA_{Hexamer}$ (H12Y)
Data collection							
Space group	$P2_1$	$C222_1$	P6	P6	P6	P6	P6
Cell dimensions							
a, *b*, *c* (Å)	43.72, 23.85, 129.55	89.69, 159.27, 249.40	90.81, 90.81, 56.68	90.73, 90,73, 56.75	90.81, 90.81, 56.88	90.76, 90.76, 56.76	90.60, 90.60, 56.93
α, β, γ (°)	90, 96.31, 90	90, 90, 90	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120	90, 90, 120
Resolution (Å)	39.87-1.76 (1.86-1.76)	19.99-3.10 (3.27-3.10)	45.98-2.03 (2.08-2.03)	26.69-1.90 (1.94-1.90)	46.09-2.00 (2.05-2.00)	45.38-1.95 (2.00-1.95)	56.93-1.70 (1.73-1.70)
R_{merge}	0.065 (0.227)	0.094 (0.552)	0.112 (0.551)	0.121 (0.748)	0.166 (0.831)	0.105 (0.862)	0.095 (0.813)
I / σI	13.3 (5.0)	7.1 (1.9)	8.2 (1.9)	8.3 (2.1)	6.5 (1.9)	11.7 (2.1)	9.0 (2.4)
Completeness (%)	91.9 (80.6)	97.4 (92.0)	94.9 (69.1)	99.8 (100.0)	100.0 (100.0)	97.9 (99.5)	95.1 (94.7)
Redundancy	4.6 (4.6)	2.4 (2.4)	2.9 (2.4)	5.1 (4.8)	6.0 (5.9)	6.8 (6.5)	4.8 (4.9)
Refinement							
Resolution (Å)	39.87-1.76 (1.81-1.76)	19.99-3.10 (3.18-3.10)	45.98-2.04 (2.10-2.04)	26.69-1.90 (1.95-1.90)	39.32-2.00 (2.05-2.00)	45.38-1.95 (2.00-1.95)	56.93-1.70 (1.74-1.70)
No. reflections	23837 (1552)	30312 (2044)	15488 (1084)	19999 (1468)	17285 (1246)	18188 (1347)	26370 (1949)
R_{work} / R_{free}	0.190/0.224 (0.233/ 0.286)	0.250/0.281 (0.368/ 0.399)	0.236/0.263 (0.316/ 0.384)	0.200/0.225 (0.266/ 0.342)	0.205/0.221 (0.216/ 0.255)	0.194/0.221 (0.266/ 0.299)	0.197/0.232 (0.265/ 0.285)
No. atoms							
Protein	2284	9358	1558	1623	1605	1612	1688
Ligand/ion	1	193	30	-	-	24	-
Water	338	-	82	124	105	119	123
B-factors							
Protein	27.928	54.173	26.577	30.265	27.732	30.878	27.523
Ligand/ion	38.516	83.117	109.040	-	-	95.409	-
Water	24.730	-	30.031	35.719	28.567	34.088	33.604
R.m.s. deviations							
Bond lengths (Å)	0.007	0.008	0.007	0.007	0.007	0.008	0.006
Bond angles (°)	1.222	1.156	0.991	0.998	1.046	1.281	1.009

Statistics for the highest resolution shell are shown in parentheses.

doi:10.1038/nature19095

Structure of mammalian respiratory complex I

Jiapeng Zhu[1]†*, Kutti R. Vinothkumar[2]* & Judy Hirst[1]

Complex I (NADH:ubiquinone oxidoreductase), one of the largest membrane-bound enzymes in the cell, powers ATP synthesis in mammalian mitochondria by using the reducing potential of NADH to drive protons across the inner mitochondrial membrane. Mammalian complex I (ref. 1) contains 45 subunits, comprising 14 core subunits that house the catalytic machinery (and are conserved from bacteria to humans) and a mammalian-specific cohort of 31 supernumerary subunits[1,2]. Knowledge of the structures and functions of the supernumerary subunits is fragmentary. Here we describe a 4.2-Å resolution single-particle electron cryomicroscopy structure of complex I from *Bos taurus*. We have located and modelled all 45 subunits, including the 31 supernumerary subunits, to provide the entire structure of the mammalian complex. Computational sorting of the particles identified different structural classes, related by subtle domain movements, which reveal conformationally dynamic regions and match biochemical descriptions of the 'active-to-de-active' enzyme transition that occurs during hypoxia[3,4]. Our structures therefore provide a foundation for understanding complex I assembly[5] and the effects of mutations that cause clinically relevant complex I dysfunctions[6], give insights into the structural and functional roles of the supernumerary subunits and reveal new information on the mechanism and regulation of catalysis.

Using structures determined for bacterial complex I (refs 7–9) as a starting point, structures of the 14 highly conserved core subunits and their nine cofactors (a flavin mononucleotide (FMN) and eight iron–sulfur (FeS) clusters) have been determined to medium resolution in complex I from both mammals (for *Bos taurus*)[10] and yeast (*Yarrowia lipolytica*)[11]. The arrangement and structures of the 31 supernumerary subunits (constituting half the mammalian complex) are, however, far less well defined. The 5-Å resolution electron cryomicroscopy (cryoEM) structure of *B. taurus* complex I revealed the supernumerary ensemble wrapped around the core, with 14 supernumerary subunits assigned[10]. Subsequently, eight further assignments were proposed using the crystallographic structure of subcomplex Iβ (part of the membrane domain)[12]. Therefore, nine subunits remain unlocated and models for the supernumerary subunits are fragmentary. The complete structure of mammalian complex I is crucial for elucidating the roles of the supernumerary subunits in complex I function and dysfunction.

Here, we describe a cryoEM map for *B. taurus* complex I with an overall resolution of 4.16 Å (Fig. 1a and Extended Data Fig. 1), which enabled modelling of all its 45 subunits and 93% of its 8,515 residues (Extended Data Tables 1, 2). Computational sorting of the particles revealed three major classes, with overall resolutions 4.27 Å (class 1), 4.35 Å (class 2) and 5.60 Å (class 3) (Extended Data Fig. 2), for which the quality of the map in several regions was improved substantially. The different classes represent different states of the complex and analysis of each provides new insights into the mechanism of complex I catalysis. Extended Data Figures 3 and 4 present example densities and we use the class 2 map and model to describe the structure, unless indicated otherwise.

Figure 1 presents the structures and locations of all 31 supernumerary subunits in mammalian complex I (see Extended Data Table 3 for subunit–subunit interactions and additional details). The supernumerary subunits are central to the structure, stability and assembly of the complex, and some also have regulatory or independent metabolic roles.

The 18 supernumerary transmembrane helices (TMHs) (Fig. 1b) establish a cage around the core membrane domain. Three TMH-containing subunits, B9 (NDUFA3 in the nomenclature for human complex I), B16.6 (NDUFA13) and MWFE (NDUFA1), interact extensively with PGIV (NDUFA8) on the intermembrane-space (IMS) face, enclosing core subunit ND1. Subunit B14.5b (NDUFC2), bound to ND2, contains two different-length TMHs and attaches KFYI (NDUFC1) to the complex. Three TMHs that interact with ND4 are assigned to MNLL (NDUFB1), ESSS (NDUFB11) and SGDH (NDUFB8). Four TMHs, assigned to B17 (NDUFB6), AGGG (NDUFB2), B12 (NDUFB3) and ASHI (NDUFB8), are bound to ND5. The TMHs of ASHI and B15 (NDUFB4, on the side of ND4) cross the ND5 transverse helix, and the four TMHs of B14.7 (NDUFA11) appear to support ND5-TMH16 in anchoring it against ND2. Four subunits confined to the IMS (PGIV, the 15 kDa subunit (NDUFS5), PDSW (NDUFB10) and B18 (NDUFB7)) form a helix latticework (together with SGDH and B16.6) on the IMS face (Fig. 2a). PGIV, the 15 kDa subunit and B18 contain CHCH domains (pairs of helices linked by two disulfide bonds)[13] and are canonical substrates for the Mia40 oxidative-folding pathway[14]; PDSW probably contains two further disulfide bonds. These disulfide bonds form during complex I biogenesis and are probably important for enzyme stability. Thus, the supernumerary cage has evolved to become integral to the structure and stability of the membrane domain.

Subunits B14 (NDUFA6) and SDAP-α (NDUFAB1), and B22 (NDUFB9) and SDAP-β (NDUFAB1), constitute matching subdomains on the hydrophilic domain and matrix face of the membrane domain, respectively[10,12,15] (Fig. 1). SDAP-α and SDAP-β are identical to the mitochondrial acyl-carrier protein (ACP) and exhibit densities consistent with the pantetheine-4'-phosphate group that covalently attaches an acyl chain to Ser44 (refs 16, 17) (Extended Data Fig. 4e). Their ACP recognition helices interact with arginine- and lysine-rich helices in the LYR proteins B14 and B22 (Fig. 2b and Extended Data Fig. 4e) as canonical ACPs interact with the enzymes of fatty-acid biosynthesis[18]. The 42 kDa subunit (NDUFA10, Fig. 2c) contains a central α/β nucleoside kinase fold with a parallel five-strand β-sheet, plus three extensions that dock it to the matrix face of ND2. Although the active site is accessible and the key nucleoside kinase residues are present[19], no activity has been reported. The 39 kDa subunit (NDUFA9, Fig. 2d) is attached to core subunits PSST (NDUFS7) and the 30 kDa subunit (NDUFS3) in the hydrophilic arm. The N-terminal domain of the 39 kDa subunit comprises an α/β short-chain dehydrogenase/reductase fold[20] containing an NAD(P)-binding Rossmann fold with a parallel seven-strand β-sheet and density for a bound nucleotide, modelled as NADPH[21] (Extended Data Fig. 4). The separate C-terminal

[1]MRC Mitochondrial Biology Unit, Hills Road, Cambridge CB2 0XY, UK. [2]MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge CB2 0QH, UK. †Present address: Nanjing University of Chinese Medicine, Nanjing 210023, China.
*These authors contributed equally to this work.

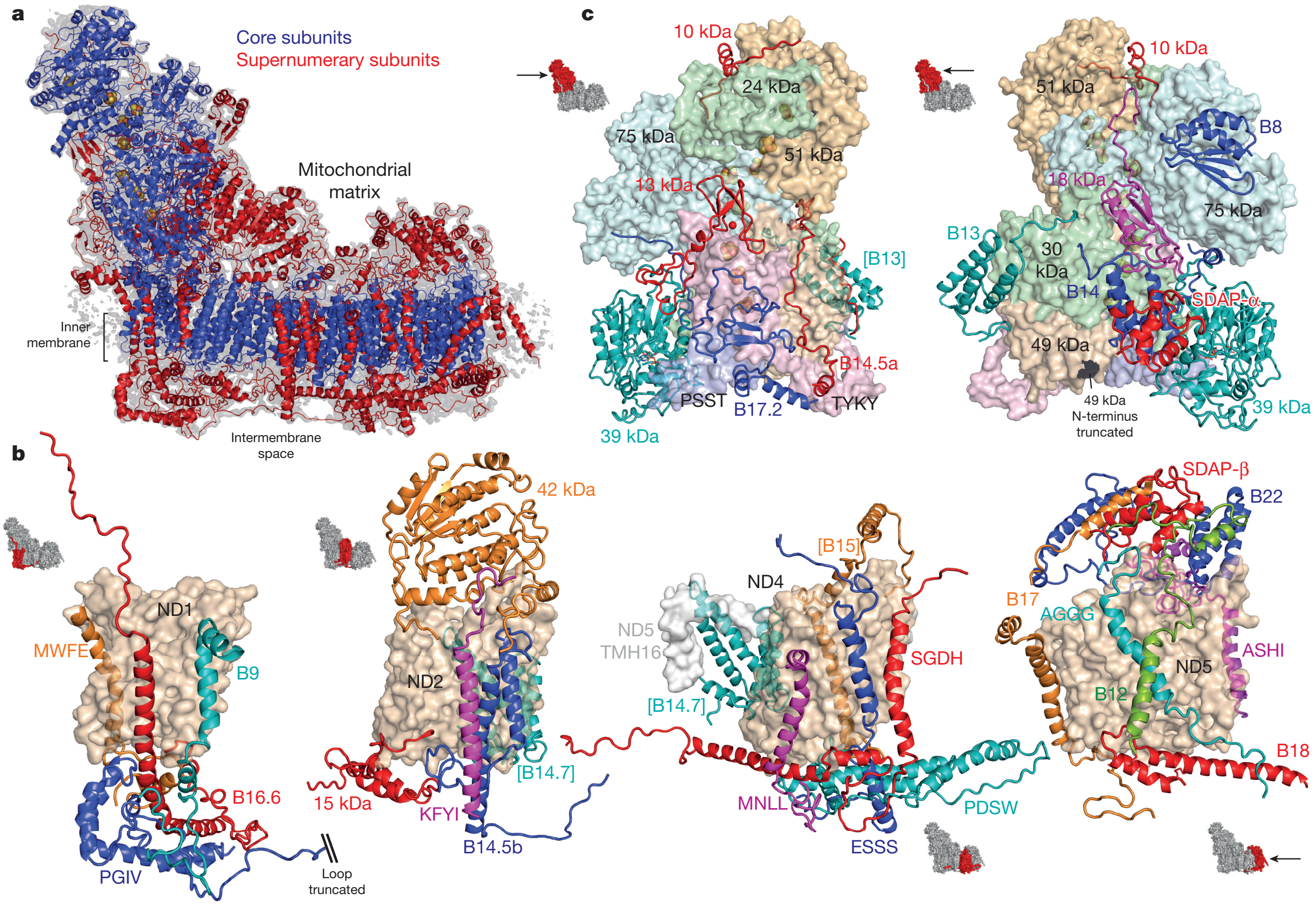


Figure 1 | The supernumerary subunits of mammalian complex I. **a**, Overview of the complex with the 14 core subunits in blue (FeS clusters in yellow/orange), the 31 supernumerary subunits in red, and the cryoEM density in grey. **b**, **c**, Arrangement and structures of the supernumerary subunits around the core membrane (**b**) and hydrophilic (**c**) subunits. The core subunits are in surface representation and the supernumerary subunits in cartoon; icons show viewpoints and locations in the complex. Subunits in brackets are behind the domain. The assignments of B12 and AGGG may be reversed. See Extended Data Tables 1 and 2 for the subunit nomenclatures in other species.

domain interacts with the long matrix loop between TMHs 1 and 2 of ND3.

The final seven supernumerary subunits adorn the hydrophilic domain (Fig. 1c). Thioredoxin-like B8 (NDUFA2) is attached to the 75 kDa subunit (NDUFS1), and the three-helix bundle of B13 (NDUFA5) to the 30 kDa subunit. The remaining five subunits are located at interfaces. The zinc-binding domain of the 13 kDa subunit (NDUFS6)[22] and the four-strand β-sheet and helix of the 18 kDa subunit (NDUFS4) are located where the NADH dehydrogenase domain meets the rest of the complex. All five subunits (the other three are B14.5a (NDUFA7), B17.2 (NDUFA12) and the 10 kDa subunit (NDUFV3)) contain long loops running over the domain surface. A notable example is the extensive loop in B14.5a, which arches up along the TYKY–49 kDa subunit (NDUFS8–NDUFS2) interface, across the 49 kDa subunit, along its interface with the 75 kDa subunit and onto the 30 kDa subunit. The role of the supernumerary subunits in stabilizing interfaces in the hydrophilic domain contrasts sharply with their arrangement into a rigid cage to stabilize the membrane domain.

The structures of the mammalian core subunits (Fig. 3a) closely match those of the bacterial subunits[7–9], and contain corresponding mechanistically relevant features. NADH is oxidized by a flavin mononucleotide in the 51 kDa subunit (NDUFV1) (Extended Data Fig. 3c). Electrons then transfer along a chain of FeS clusters to the terminal cluster (N2) and to ubiquinone-10. In mammalian complex I, an unusual dimethylated arginine (Arg85 of the 49 kDa subunit)[23] close to N2 probably contributes to its relatively high reduction potential[24]. In the hydrophilic domain, the large domain of the mammalian 75 kDa subunit differs from that of *Thermus thermophilus*[7] as its fourth sub-domain contains just two short helices separated by a loop of around 30 residues. The core subunit N- and C-terminal extensions also vary between species; in *B. taurus* the C terminus of the 30 kDa subunit (which loops into a cleft between PSST, the 39 kDa subunit and B14) and the N terminus of TYKY (which wraps around the hydrophilic/hydrophobic domain interface) recapitulate the stabilizing role of the supernumerary subunits. Notably, our cryoEM maps reveal that the 49 kDa subunit N terminus forms a long loop on the surface of the membrane domain (Fig. 3a). This extended conformation explains its susceptibility to proteases[25], but it is unlikely to be central to the mechanism because it is not conserved in *T. thermophilus*, and in *Escherichia coli* is fused to the C terminus of the 30 kDa subunit. The long matrix loop in ND3, which lies across the front of the hydrophilic domain and is central to the transition between active and de-active states in mammalian complex I (ref. 3), is also resolved (Fig. 3a).

Four proton-transfer routes (in ND2, ND4, ND5 (ref. 8), and ND1 + ND4L + ND6 (ref. 9)) have been proposed for the four protons that complex I is generally considered to translocate for each NADH molecule oxidized. ND2, ND4 and ND5 each contain two TMHs interrupted by loops in the central membrane plane (TMH4 and TMH9 in ND2; TMH7 and TMH12 in ND4 and ND5, Fig. 3a and Extended Data Fig. 3). The chain of conserved aspartate, glutamate, lysine and histidine residues that runs along the middle of the membrane domain is now well defined in the mammalian enzyme (Glu143 of ND1, Asp66 and Glu68 of ND3; Glu34 and Glu70 of ND4L; Glu34, Lys105, Lys135

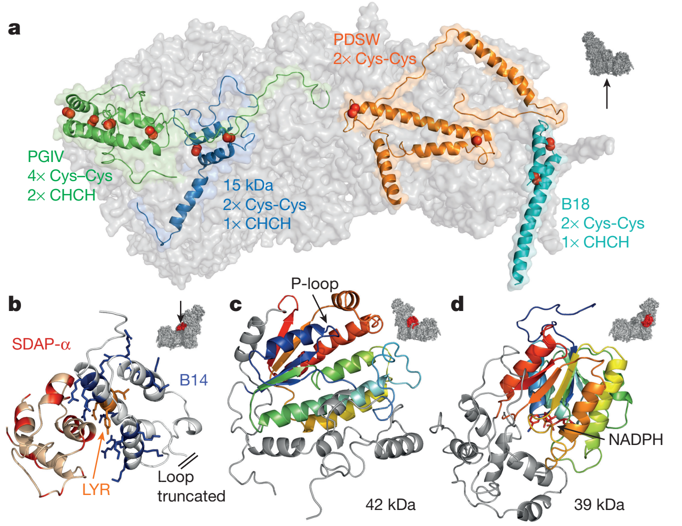


Figure 2 | Details of some of the supernumerary subunits. a, Subunits confined to the IMS face with disulfide bonds indicated by red spheres. **b**, Positively charged residues in the LYR-protein B14 (blue) interact with negatively charged residues in SDAP-α (red); B22 and SDAP-β exhibit the same structural motif. **c**, The 42 kDa subunit with the core nucleoside-kinase fold in rainbow; the extensions (grey) dock it to ND2. **d**, The 39 kDa subunit with the core dehydrogenase-reductase fold in rainbow and bound nucleotide; the C-terminal domain (grey) approaches the membrane interface. Icons indicate viewpoints and locations in the complex.

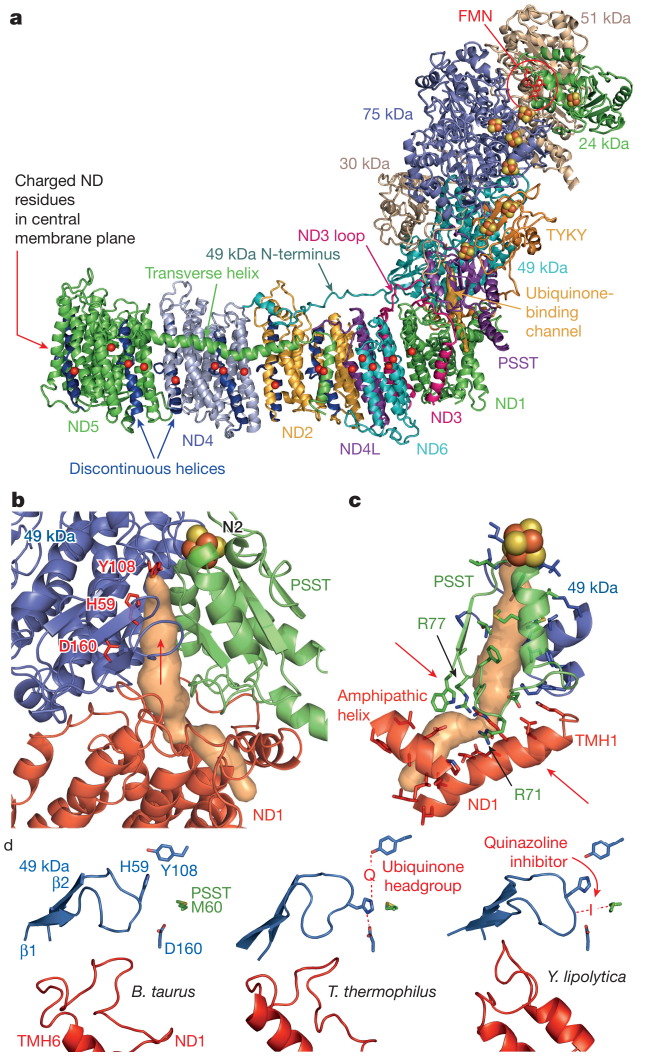


Figure 3 | The core subunits and ubiquinone-binding site of mammalian complex I. a, The core subunits with the FMN, FeS clusters, conserved charged residues (Cα) in the membrane (overlaid for clarity), discontinuous TMHs (blue), and proposed ubiquinone-binding channel (orange). **b**, The ubiquinone-binding channel. Tyr108 and His59 of the 49 kDa subunit[9] form hydrogen bonds to the bound ubiquinone at the top of the cleft (indicated by an arrow) between the 49 kDa subunit and PSST. The channel entrance is between three helices in ND1. **c**, Structural elements forming the channel and bottleneck (between the arrows). Arg77 of PSST is hydroxylated[23]. **d**, Conformations of the 49 kDa subunit (β1–β2) and ND1 (TMH5–6) loops observed in different species. In *T. thermophilus*[9] the ubiquinone headgroup binds between Tyr108 and His59 and His59 hydrogen bonds to Asp160 of the 49 kDa subunit. In *Y. lipolytica*[11] a quinazoline inhibitor is bound between Met60 of PSST and the tip of the 49 kDa subunit loop.

and Lys263 of ND2; Glu123, Lys206, Lys237 and Glu378 of ND4; Glu145, Lys223, His248 and Lys392 of ND5, Fig. 3a). Distortions of the helical structure are observed in TMH3 of ND6, TMH5 of ND2, TMH8 of ND4 and TMH8 of ND5 (Extended Data Fig. 3). These distortions resemble the π–bulge in bacteriorhodopsin but do not satisfy its technical definition[26], perhaps owing to the intermediate resolution of the maps. The distortions are centred on glycine pairs in ND6 (62–3) and ND4 (239–40), on a serine pair in ND5 (249–50), and on Trp167 (flanked by two glycine pairs) in ND2. Notably, TMH3 of ND6 is more distorted in the mammalian structure than in *T. thermophilus* (which contains only one glycine residue)[9], such that Phe67 of ND6 is displaced around the helical axis.

Ubiquinone-10 binds with its redox-active headgroup close to cluster N2, at the top of a cleft between the 49 kDa subunit and PSST[7,27] (Fig. 3b), while *T. thermophilus* complex I co-crystallized with decylubiquinone showed it forms hydrogen bonds with His59 and Tyr108 of the 49 kDa subunit[9]. Here, the side chains of Tyr108 and His59 are poorly resolved, and the conformation of the β1–β2 His59-containing loop is different to that in *T. thermophilus* (Fig. 3d). It therefore appears that the structural elements that form the binding site are flexible, allowing it to organize around substrates and inhibitors (neither of which are present here). The putative ubiquinone-access channel, identified first in *T. thermophilus*[9], connects the cleft to an entrance in ND1 (between TMH1, an amphipathic helix, and TMH6) and can also be detected here (minimum diameter, 2.9 Å). Alternative entrances, between TMH1 and TMH7, and TMH5 and TMH6 of ND1, are also evident but narrower (minimum diameters, 1.9—2.2 Å). However, the planar ubiquinone ring is approximately 6 Å across, so all the channels in the static structure would have to open to allow it to enter. A structure containing ubiquinone-10 (or a long-chain analogue) is therefore required to confirm its access pathway.

In the mammalian complex, further consideration of the most plausible channel (that is, the widest) for ubiquinone reveals a 'bottleneck' at the base of the cleft (Fig. 3c). Ubiquinone-10 is highly hydrophobic so most of the channel-lining residues are uncharged and hydrophobic. In contrast, the bottleneck is formed by charged and polar residues including Glu24 and Arg25 (TMH1 of ND1), Arg274 (TMH7 of ND1), and Arg71 and hydroxy-Arg77 (ref. 23) (α2–β2 loop of PSST). Nearby, the TMH5–6 loop of ND1, with many acidic residues, contributes more notably to channel formation in *T. thermophilus* and *Y. lipolytica*[9,11]. This cluster of charged residues suggests the presence of water molecules and appears to be incompatible with a ubiquinone-10 binding channel. However, the PSST loop was modelled incompletely in *T. thermophilus* and *Y. lipolytica*, and the ND1 loop is poorly resolved here, indicating their flexibility. It is possible that conformational changes at the bottleneck, linked to ubiquinone binding and dissociation, contribute to coupling of the redox reaction to proton translocation.

When the particles comprising the whole data set were subjected to 3D classification, three major, slightly different classes emerged. Class 3,

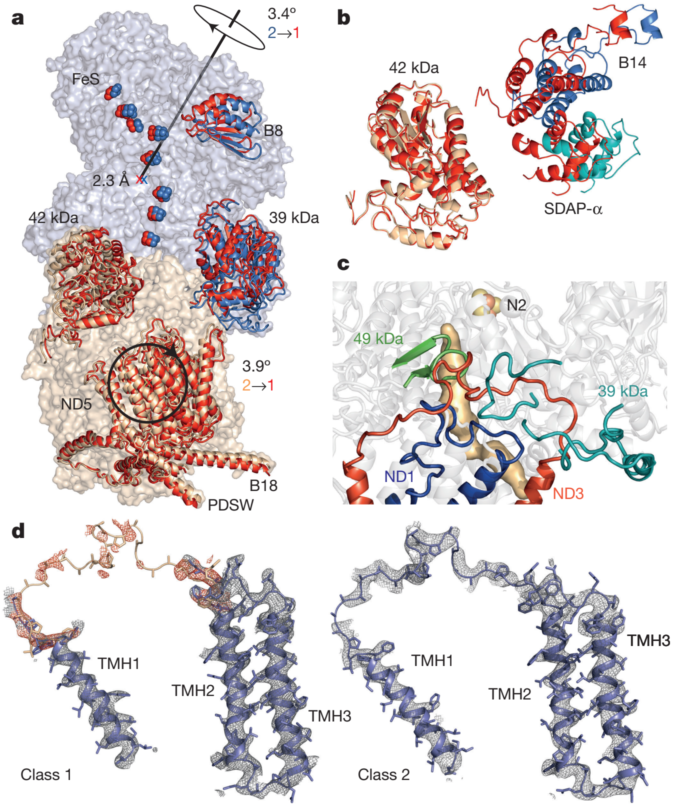


Figure 4 | Relationships between classes 1 and 2. a, Class 1 (red) and 2 (blue/wheat) were superimposed using ND1 and ND3 and viewed along the axis of rotation for ND4 and ND5. See Extended Data Table 4 for details of the transformations. **b**, Change of approximately 10 Å in the relative positions of B14 and SDAPα (hydrophilic domain) and the 42 kDa subunit (membrane domain); class 1 in red, viewed from the matrix. **c**, Loops (ND1 TMH5–6, ND3 TMH1–2, β1–β2 of the 49 kDa subunit and in the 39 kDa subunit) in class 2 that are disordered in class 1, with the ubiquinone-binding site; adjacent TMHs and strands are shown for clarity. The site cannot be detected in class 1 as it appears open. **d**, Densities for the loop connecting TMHs 1 and 2 of ND3. For class 1, the loop from the class 2 model (white) was used to identify the density in red.

the smallest, lowest resolution class, is closer to class 1 than class 2 in structure and is characterized by movement of the ND4–ND5 subdomain (relative to class 1, Extended Data Fig. 5 and Extended Data Table 4) and disorder in the ND5 transverse helix and its anchor (TMH16 of ND5). Similar disorder was observed in subcomplex Iβ (ref. 12), which comprises the ND4–ND5 subdomain. We therefore suggest that class 3 is a state in which molecules are in the first stages of dissociation and do not discuss it further.

Classes 1 and 2 are related (Fig. 4a) by opposing rotations of the hydrophilic domain and a large section of the membrane domain, relative to the ND1 subdomain (Extended Data Table 4). In class 1, the 42 kDa subunit has moved towards B14 and SDAPα (Fig. 4b), and the 39 kDa subunit has moved relative to ND1. Notably, the long TMH1–2 loop of ND3 is partially disordered (Fig. 4d). This loop is symptomatic of decreased order in class 1 at the hydrophilic-membrane domain junction; the TMH5–6 loop of ND1, the β1–β2 loop of the 49 kDa subunit containing His59, and parts of the C-terminal domain of the 39 kDa subunit are also disordered (Fig. 4c). In addition, the distortion in TMH3 of ND6 is less pronounced in class 1 than class 2.

Mammalian complex I exists in different states according to its catalytic status. In the absence of substrates to sustain turnover (such as during hypoxia) it converts spontaneously to its 'de-active' state[4], a profound resting state that requires slow, reactivating turnovers to regain 'active' status. The de-active state is characterized by the ability of cysteine-modifying reagents (such as *N*-ethylmaleimide) to derivatize Cys39 in the ND3 TMH1–2 loop[3]. Approximately half the preparation discussed here is susceptible to modification by *N*-ethylmaleimide. In class 2, the side chain of Cys39 of ND3 is inaccessible to modifying reagents, suggesting class 2 represents an active state of the complex. By contrast, the disordered loop in class 1 (Fig. 4d) is probably mobile and accessible, suggesting class 1 represents a de-active state. Increased disorder in the C-terminal domain of the 39 kDa subunit, and its altered position relative to ND1, support this assignment because both subunits are more exposed to lysine-modifying reagents in the de-active state[28]. However, the structures of biochemically defined samples are required to confirm these assignments.

Different conformations of the ND1 TMH5–6 loop and the β1–β2 loop of the 49 kDa subunit in *Y. lipolytica* (relative to that in *T. thermophilus*) were proposed previously as characteristic of the de-active state[11], but they vary between our class 2 conformation and that of *T. thermophilus* (Fig. 3d), and are disordered in class 1. Notably, *Y. lipolytica* complex I was co-crystallized with a quinazoline inhibitor (Fig. 3d), and cross-linking studies have shown quinazolines interact with sections of the 49 kDa subunit and ND1 that contain the β1–β2 and TMH5–6 loops[25]. We propose quinazoline binding orders these loops, and the quinazoline-binding site overlaps with (but does not superimpose on) the ubiquinone-binding site. Our interpretation supports biochemical proposals for non-identical but overlapping sites for the myriad inhibitors of ubiquinone reduction[29], but does not support an alternative, occluded ubiquinone-binding site in the de-active complex[11].

The two states of mammalian complex I described support the idea that dynamic, flexible regions at the hydrophilic–membrane domain interface are important for coupling ubiquinone reduction to proton translocation. The class-1-disordered loops in ND1, ND3 and the 49 kDa subunit all contribute to the ubiquinone-binding site (Fig. 4c). Therefore, we attribute lack of catalytic activity in the de-active state to reversible disruption of this site, which can be recovered when the ubiquinone-binding site in the NADH-reduced enzyme reforms around its substrate. During catalysis, the ND3 loop, which originates in the membrane and interacts extensively with the hydrophilic domain, may restrict conformational changes at the domain interface. Changes in the conformation of the loop of the 49 kDa subunit may trigger proton translocation: molecular simulations were used to outline a mechanism in which the ubiquinol dianion deprotonates Tyr108 and His59, breaking a His59–Asp160 hydrogen bond and displacing Asp160 towards the membrane[30]. In ND1, TMH2–6 replicate the antiporter-like half-channel motif of ND2, ND4 and ND5 (ref. 9). TMH5 resembles a discontinuous TMH, but with its half helix on the matrix side unstructured and continuous with the TMH5–6 loop at the base of the ubiquinone-binding cleft (Fig. 3d). Like α2–β2 in PSST (Fig. 3c), this loop may change conformation upon ubiquinone binding. Furthermore, the loop carries many conserved acidic residues that may collect protons for Glu143 of ND1 (ref. 9). In turn, Glu143 is connected to the chain of charged residues along the membrane domain by Asp66 of ND3 and the dynamic distortion in TMH3 of ND6 (Fig. 3a). Thus, a cascade of events originating from the ubiquinone-binding cleft may couple ubiquinone reduction and protonation to proton translocation. Although all such mechanisms for complex I are currently hypothetical, cryoEM now provides a powerful tool to study individual trapped conformations or separate mixed states computationally in order to determine how conformational changes are initiated, coordinated and propagated.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 11 March; accepted 8 July 2016.
Published online 10 August 2016.

1. Hirst, J. Mitochondrial complex I. *Annu. Rev. Biochem.* **82,** 551–575 (2013).
2. Hirst, J., Carroll, J., Fearnley, I. M., Shannon, R. J. & Walker, J. E. The nuclear encoded subunits of complex I from bovine heart mitochondria. *Biochim. Biophys. Acta* **1604,** 135–150 (2003).
3. Galkin, A. *et al.* Identification of the mitochondrial ND3 subunit as a structural component involved in the active/deactive enzyme transition of respiratory complex I. *J. Biol. Chem.* **283,** 20907–20913 (2008).
4. Galkin, A., Abramov, A. Y., Frakich, N., Duchen, M. R. & Moncada, S. Lack of oxygen deactivates mitochondrial complex I: implications for ischemic injury? *J. Biol. Chem.* **284,** 36055–36061 (2009).
5. Sánchez-Caballero, L., Guerrero-Castillo, S. & Nijtmans, L. Unraveling the complexity of mitochondrial complex I assembly: A dynamic process. *Biochim. Biophys. Acta* **1857,** 980–990 (2016).
6. Fassone, E. & Rahman, S. Complex I deficiency: clinical features, biochemistry and molecular genetics. *J. Med. Genet.* **49,** 578–590 (2012).
7. Sazanov, L. A. & Hinchliffe, P. Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*. *Science* **311,** 1430–1436 (2006).
8. Efremov, R. G. & Sazanov, L. A. Structure of the membrane domain of respiratory complex I. *Nature* **476,** 414–420 (2011).
9. Baradaran, R., Berrisford, J. M., Minhas, G. S. & Sazanov, L. A. Crystal structure of the entire respiratory complex I. *Nature* **494,** 443–448 (2013).
10. Vinothkumar, K. R., Zhu, J. & Hirst, J. Architecture of mammalian respiratory complex I. *Nature* **515,** 80–84 (2014).
11. Zickermann, V. *et al.* Structural biology. Mechanistic insight from the crystal structure of mitochondrial complex I. *Science* **347,** 44–49 (2015).
12. Zhu, J. *et al.* Structure of subcomplex Iβ of mammalian respiratory complex I leads to new supernumerary subunit assignments. *Proc. Natl Acad. Sci. USA* **112,** 12087–12092 (2015).
13. Szklarczyk, R. *et al.* NDUFB7 and NDUFA8 are located at the intermembrane surface of complex I. *FEBS Lett.* **585,** 737–743 (2011).
14. Banci, L. *et al.* MIA40 is an oxidoreductase that catalyzes oxidative protein folding in mitochondria. *Nat. Struct. Mol. Biol.* **16,** 198–206 (2009).
15. Angerer, H. *et al.* The LYR protein subunit NB4M/NDUFA6 of mitochondrial complex I anchors an acyl carrier protein and is essential for catalytic activity. *Proc. Natl Acad. Sci. USA* **111,** 5207–5212 (2014).
16. Runswick, M. J., Fearnley, I. M., Skehel, J. M. & Walker, J. E. Presence of an acyl carrier protein in NADH:ubiquinone oxidoreductase from bovine heart mitochondria. *FEBS Lett.* **286,** 121–124 (1991).
17. Feng, D., Witkowski, A. & Smith, S. Down-regulation of mitochondrial acyl carrier protein in mammalian cells compromises protein lipoylation and respiratory complex I and results in cell death. *J. Biol. Chem.* **284,** 11436–11445 (2009).
18. Chan, D. I. & Vogel, H. J. Current understanding of fatty acid biosynthesis and the acyl carrier protein. *Biochem. J.* **430,** 1–19 (2010).
19. Johansson, K. *et al.* Structural basis for substrate specificities of cellular deoxyribonucleoside kinases. *Nat. Struct. Biol.* **8,** 616–620 (2001).
20. Fearnley, I. M. & Walker, J. E. Conservation of sequences of subunits of mitochondrial complex I and their relationships with other proteins. *Biochim. Biophys. Acta* **1140,** 105–134 (1992).
21. Abdrakhmanova, A., Zwicker, K., Kerscher, S., Zickermann, V. & Brandt, U. Tight binding of NADPH to the 39-kDa subunit of complex I is not required for catalytic activity but stabilizes the multiprotein complex. *Biochim. Biophys. Acta* **1757,** 1676–1682 (2006).
22. Kmita, K. *et al.* Accessory NUMM (NDUFS6) subunit harbors a Zn-binding site and is essential for biogenesis of mitochondrial complex I. *Proc. Natl Acad. Sci. USA* **112,** 5685–5690 (2015).
23. Carroll, J., Ding, S., Fearnley, I. M. & Walker, J. E. Post-translational modifications near the quinone binding site of mammalian complex I. *J. Biol. Chem.* **288,** 24799–24808 (2013).
24. Hirst, J. & Roessler, M. M. Energy conversion, redox catalysis and generation of reactive oxygen species by respiratory complex I. *Biochim. Biophys. Acta* **1857,** 872–883 (2016).
25. Murai, M., Mashimo, Y., Hirst, J. & Miyoshi, H. Exploring interactions between the 49 kDa and ND1 subunits in mitochondrial NADH-ubiquinone oxidoreductase (complex I) by photoaffinity labeling. *Biochemistry* **50,** 6901–6908 (2011).
26. Cooley, R. B., Arp, D. J. & Karplus, P. A. Evolutionary origin of a secondary structure: π-helices as cryptic but widespread insertional variations of α-helices that enhance protein functionality. *J. Mol. Biol.* **404,** 232–246 (2010).
27. Tocilescu, M. A., Fendel, U., Zwicker, K., Kerscher, S. & Brandt, U. Exploring the ubiquinone binding cavity of respiratory complex I. *J. Biol. Chem.* **282,** 29514–29520 (2007).
28. Babot, M. *et al.* ND3, ND1 and 39kDa subunits are more exposed in the de-active form of bovine mitochondrial complex I. *Biochim. Biophys. Acta* **1837,** 929–939 (2014).
29. Murai, M. & Miyoshi, H. Current topics on inhibitors of respiratory complex I. *Biochim. Biophys. Acta* **1857,** 884–891 (2016).
30. Sharma, V. *et al.* Redox-induced activation of the proton pump in the respiratory complex I. *Proc. Natl Acad. Sci. USA* **112,** 11571–11576 (2015).

Acknowledgements We thank R. Henderson and G. Murshudov for discussions, J. Grimmett and T. Darling for computational help, S. Chen and C. Savva for electron microscopy help, and J. N. Blaza for quantifying the ratio of de-active and active enzymes in our preparation. This work was supported by The Medical Research Council, grant numbers U105184322 (K.R.V., in R. Henderson's group) and U105663141 (J.H.).

Author Contributions J.Z. prepared protein; K.R.V. carried out electron microscopy data collection and analysis with help from J.Z.; J.Z. built the initial model; J.Z., K.R.V. and J.H. worked together, led by J.H., to model and analyse the data; J.H. designed the project; J.H. wrote the paper with help from J.Z. and K.R.V.

Author Information The electron microscopy maps and models for each class have been deposited in the Electron Microscopy Data Bank (EMDB) with accession numbers EMD-4040 (class 1), EMD-4032 (class 2) and EMD-4041 (class 3), and in the Protein Data Bank with accessions 5LDW (class 1), 5LC5 (class 2) and 5LDX (class 3). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to K.R.V. (vkumar@mrc-lmb.cam.ac.uk) or J.H. (jh@mrc-mbu.cam.ac.uk).

Reviewer Information *Nature* thanks R. B. Gennis, M. T. Ryan and the other anonymous reviewer(s) for their contribution to the peer review of this work.

METHODS

Protein preparation. Complex I was purified from *B. taurus* heart mitochondrial membranes by solubilization and anion-exchange chromatography in *n*-dodecyl-β-D-maltoside (DDM), and size-exclusion chromatography in 7-cyclohexyl-1-heptyl-β-D-maltoside (Cymal 7), as described previously[10,31].

CryoEM specimen preparation, imaging and image processing. CryoEM grids were prepared as described previously[10]. Images were recorded using a 300 keV FEI Titan Krios electron microscope with EPU software, with the specimen temperature at −186 °C. A Falcon II CMOS (complementary metal oxide semiconductor) direct electron detector was used for imaging at 105,263× magnification (nominally 59,000×), corresponding to a sampling density of 1.33 Å pixel^{-1}, at 2.4–4.2 μm under-focus in 0.3 μm steps, with the autofocus routine performed every 8 μm to give a range of defocus. Each image was exposed for 2 s (total dose ~35 e$^-$ Å^{-2}) and an in-house protocol was used to capture 34 movie frames. The frames were aligned using Unblur (without dose filtering)[32] and the CTF was determined with CTFFIND4 (ref. 33). A total of 139,456 particles were picked manually and extracted using a box of 360 pixels. Further processing was performed with RELION[34]. Following an initial 2D classification to discard 'bad' particles, 115,966 particles were used for refinement. The 5 Å resolution map described previously[10] was low-pass filtered to 60 Å and used as the reference map. The whole data set was subjected to the auto-refine routine in RELION, followed by modelling of the beam-induced movement (using a running average of 7 frames) and B-factor weighting. All the resolutions described here are defined at FSC = 0.143 following application of a shape mask, phase randomization to check for effects of the mask, and correction for the modulation transfer function of the detector. The resolution of the data set containing all the particles after B-factor weighting and refinement was 4.16 Å, with an estimated orientation accuracy of 0.93 degrees.

The B-factor weighted particles were subjected to 3D classification into eight classes using an angular sampling of 0.9° for 25 iterations, with the resolution limited to 8 Å. Three major classes were identified, containing 48,033 (class 1), 33,301 (class 2) and 19,306 (class 3) particles. Each class was refined individually, providing maps with overall resolution of 4.27 Å for class 1, 4.35 Å for class 2 and 5.6 Å for class 3. The maps were sharpened with B-factors of -114 for class 1, -110 for class 2 and -125 for class 3. Local analysis of the resolution was performed using ResMap[35] (Extended Data Figs 1, 2). Note that the map used as the reference for refinement is a class 1 map, which we described previously as the major class[10], and that the number of particles required to achieve the reported resolutions indicates the need for future improvements in both the biochemical homogeneity and specimen preparation of the samples.

Model building and analysis. Model building was performed using Coot[36]. The first model was built using the map from the complete data set, with cross-referencing to the maps from classes 1 and 2, using the 5 Å model for *B. taurus* complex I described previously (PDB accession code 4UQ8 (ref. 10)) as the initial template. This unrefined polyalanine model contains models for the fourteen core subunits that are structural homologues of the subunits of the bacterial enzyme[9], partial models for fourteen assigned supernumerary subunits, and a further 21 polypeptide chains from unassigned supernumerary structures. Assignments to some of these chains were subsequently proposed for a further eight supernumerary subunits using the 6.8 Å X-ray crystallographic structure of subcomplex Iβ from *B. taurus* complex I (ref. 12). The new maps show clear connectivity within the density features, allowing many of the previously traced chains to be extended considerably, and some of them to be joined together. Furthermore, the helical pitches of most of the TMHs and of many of the helices in the globular subunits are now clear, and the β-strands are well separated. These substantial improvements in the density, together with information from secondary structure analyses and homologous structures, allowed improved and more complete models to be built for all 45 subunits. Note that the former subunit MLRQ (NDUFA4) is no longer considered a subunit of complex I[37] and that there are two copies of subunit SDAP[10]. The only substantial un-modelled protein densities are underneath the tip of the membrane domain and are accounted for by the termini of two supernumerary subunits (B18 and ASHI).

In well-resolved regions of the map, protruding densities of the side chains of the bulky aromatic residues Phe, Tyr and Trp, along with some side chain densities from Arg and His, are clearly visible (Extended Data Figs 3, 4). For those subunits that had already been assigned, these side chain features were used as landmarks for assigning the sequences. Side chains were added in well-resolved regions, but omitted when their density features are unclear. The assignments of four subunits that were previously assigned in pairs (B9 and MWFE[10], and PDSW and B18 (ref. 12)) were also confirmed. For three subunits (B8, SDAP-α and SDAP-β) models of the human homologues are available in the PDB (PDB accession codes 1S3A (ref. 38) and 2DNW) and were used to assign the residues. For highly conserved regions of the 51 kDa, 24 kDa (NDUFV2) and 75 kDa subunits, residue assignments were supported by the structure of complex I from *T. thermophilus*[9]. In some less well-resolved regions of the map it is not possible to assign the sequence confidently using the current data. In these regions the polyalanine model has been retained (residue names UNK), but the residues for each subunit have been numbered as accurately as possible, to provide a guide to the location of individual residues. The exceptions are subunits B12, 10 kDa and part of B14.5a, for which residue numbers cannot be confidently proposed. In summary, the models described are mixed models in which the residues in some subunits are fully assigned, some partly assigned, and others not assigned at all (Extended Data Tables 1, 2). Next, the hitherto-unknown subunits were assigned. The patterns of the bulky residues observed in the TMH-containing regions of unknown densities were compared with the amino acid sequences of candidate subunits. By combining this information with information from secondary structure analyses, supported with biochemical knowledge, the seven hitherto-unassigned TMHs were assigned and the subunits modelled as described above. Subunits B12 and AGGG were assigned to the two TMHs on the tip of the membrane domain, but lack of clear features in the densities means that our specific assignment of B12 to chain k and AGGG to chain j is less confident, so we cannot exclude the possibility that they have been reversed. The two additional TMH-like densities observed in the structure of subcomplex Iβ (ref. 12) are clearly absent from the cryoEM maps, so they are attributed to an artefact produced by crystal contacts with either dissociated subunits or the reorganized transverse helix. Remaining polypeptide chains, located on the outside of the hydrophilic domain and the IMS face of the membrane domain, were assigned by combining secondary structure analyses with biochemical knowledge and by using the densities from bulky residues, and the residues and side chains assigned and modelled where possible. Maps with different B-factor sharpening were used to help with chain tracing and assignment of residues, and the model was checked for consistency with the individual maps from the three different classes. The geometry of the model was improved by cycles of manual adjustment in Coot[36], real-space refinement by Phenix[39] and refinement by REFMAC[40], with secondary structure restraints.

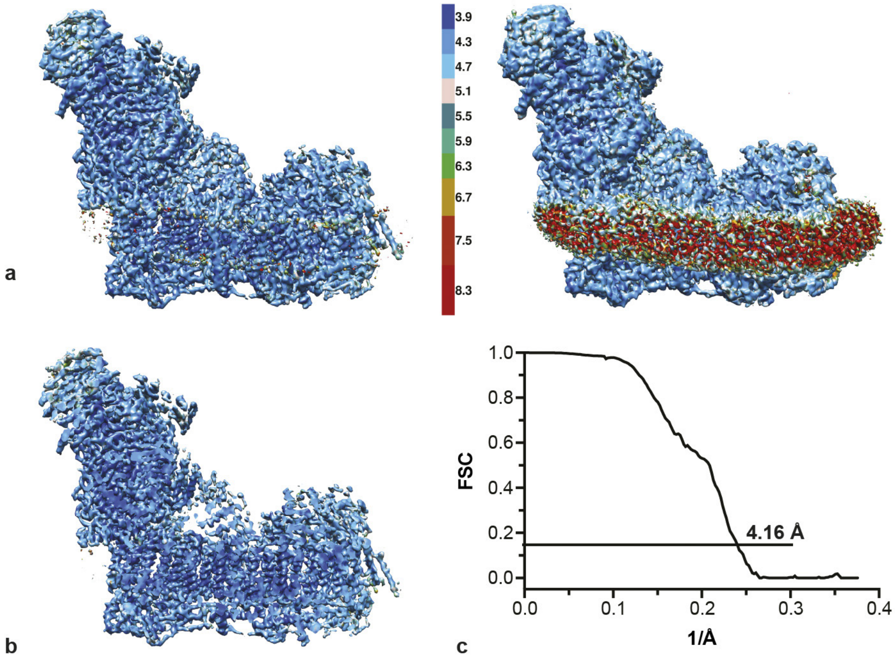
Separate models were subsequently created for classes 1 and 2 by rigid-body fitting of each subunit, manual adjustment to account for substantial local differences identified in the density maps, and cycles of adjustment and refinement as described above. To provide some reassurance of the refinement, the coordinates for classes 1 and 2 were shifted by 0.1 Å and the B-factors re-set to 75 and the coordinates were then refined against one of the half maps. The resulting models were then used to calculate Fourier-shell correlation (FSC) curves for both half maps; as shown in Extended Data Fig. 2 the curves display very little evidence of over-fitting. The class 3 model was created from the class 1 model by rigid-body fitting. It was created purely for comparison with classes 1 and 2 so the individual subunit models were retained unchanged from class 1, the model was not deleted in the regions of poor density reported in Extended Data Fig. 5, and no further refinement was performed. The refinement statistics for classes 1 and 2 are summarized in Extended Data Table 5.

The sequences of all the subunits are numbered starting from residue 1 of the mature proteins[2]. The naming of the chains has been retained as much as possible from our previous model (PDB accession code 4UQ8). The names are unchanged for chains A–Z and a–n, except that the previous chains d and e have been combined to form new chain d (B14.5b) and chain e has been reallocated to the 15 kDa subunit. Previous chains o–w all represented sections of subunits that have now been combined; new chains n–s have been reallocated.

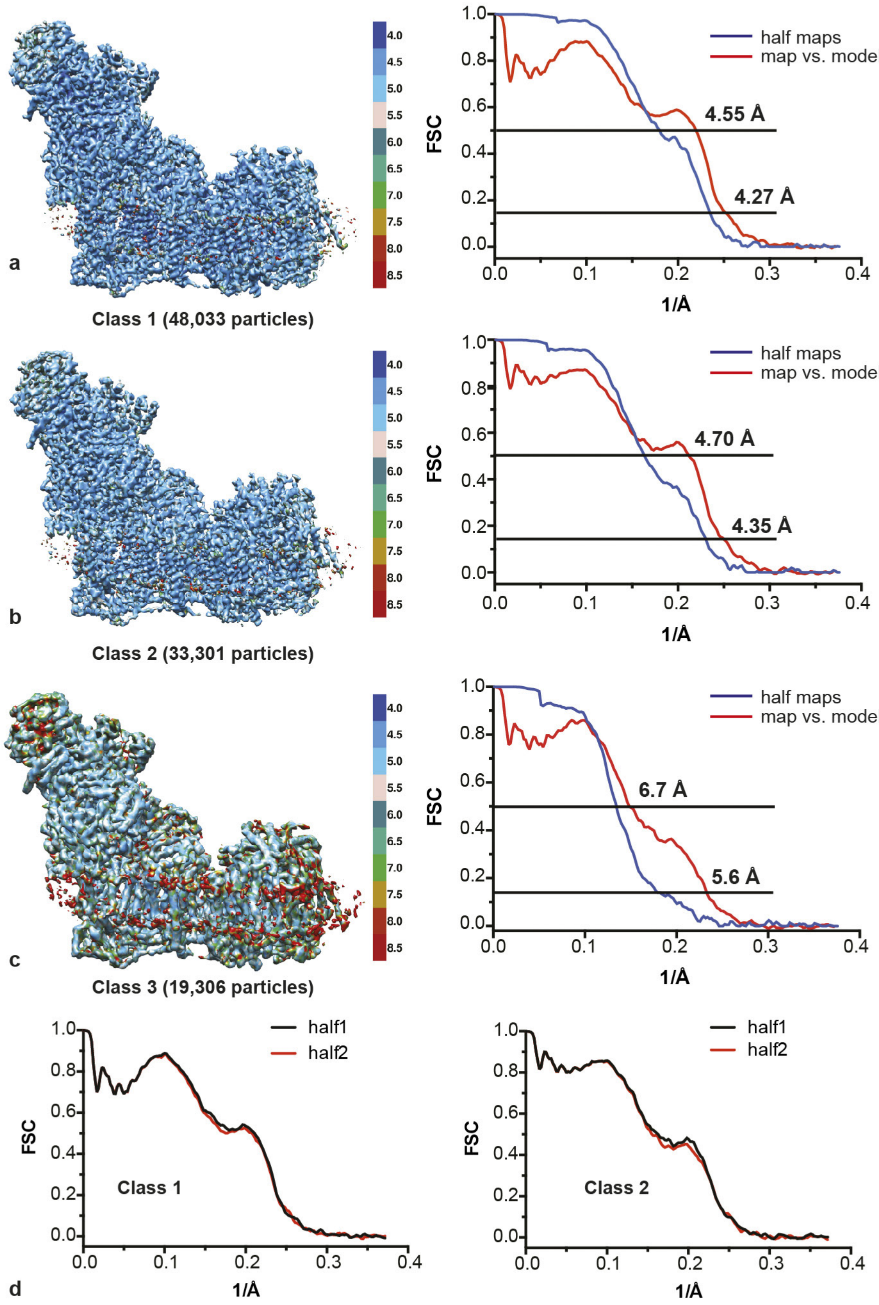
After modelling the protein, we observed two additional short, elongated densities located at interfaces between core membrane-domain subunits, which may represent phospholipid molecules. They are between TMH10 of ND2 (residues 291–295), TMH5 of ND4 (residues 144–147) and the ND5 transverse helix (residues 564–567), and between TMH11 of ND4 (residues 356–360) and TMH4 of ND5 (residues 116–123). These densities have not been modelled because similar densities are observed elsewhere within the detergent/phospholipid belt but at lower contour levels. As such, we cannot exclude the possibility that they are due to noise in the density map at the current resolution.

Bioinformatics. Secondary structure analyses were carried out using PSIPRED[41] and raptorX[42]. The identification of TMHs in the sequences and the structures of homologous proteins were described previously[10]. Cavities and channels in the structures were investigated using CAVER[43]. Figures were created using the PyMOL Molecular Graphics System. The subunit interactions in Extended Data Table 3 were calculated with NCONT in CCP4 (ref. 44), and defined as a centre-to-centre distance of less than 5 Å between any two atoms. For some subunits, such as B8 and SDAP-α, the interactions are limited to one subunit, while other subunits with long loops and extended structures, such as SGDH, form multiple interactions. Some of the residues in our current model lack side chains so the number of interactions detected may increase in future models.

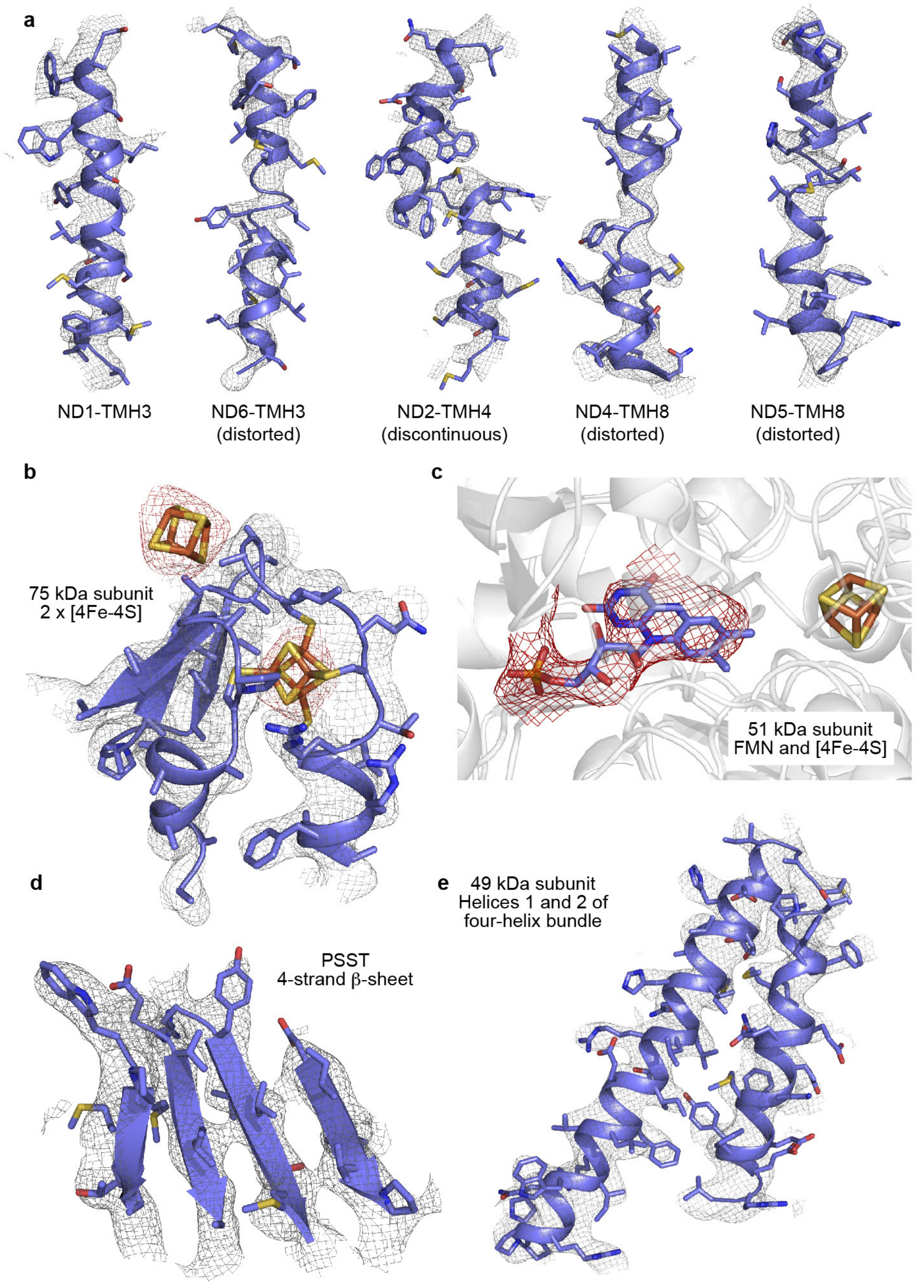
31. Sharpley, M. S., Shannon, R. J., Draghi, F. & Hirst, J. Interactions between phospholipids and NADH:ubiquinone oxidoreductase (complex I) from bovine mitochondria. *Biochemistry* **45,** 241–248 (2006).
32. Grant, T. & Grigorieff, N. Measuring the optimal exposure for single particle cryo-EM using a 2.6 Å reconstruction of rotavirus VP6. *eLife* **4,** e06980 (2015).
33. Rohou, A. & Grigorieff, N. CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192,** 216–221 (2015).
34. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180,** 519–530 (2012).
35. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11,** 63–65 (2014).
36. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of *Coot*. *Acta Crystallogr. D* **66,** 486–501 (2010).
37. Balsa, E. *et al.* NDUFA4 is a subunit of complex IV of the mammalian electron transport chain. *Cell Metab.* **16,** 378–386 (2012).
38. Brockmann, C. *et al.* The oxidized subunit B8 from human complex I adopts a thioredoxin fold. *Structure* **12,** 1645–1654 (2004).
39. Adams, P. D. *et al. PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66,** 213–221 (2010).
40. Brown, A. *et al.* Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71,** 136–153 (2015).
41. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292,** 195–202 (1999).
42. Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat. Protocols* **7,** 1511–1522 (2012).
43. Chovancova, E. *et al.* CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLOS Comput. Biol.* **8,** e1002708 (2012).
44. Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D* **50,** 760–763 (1994).



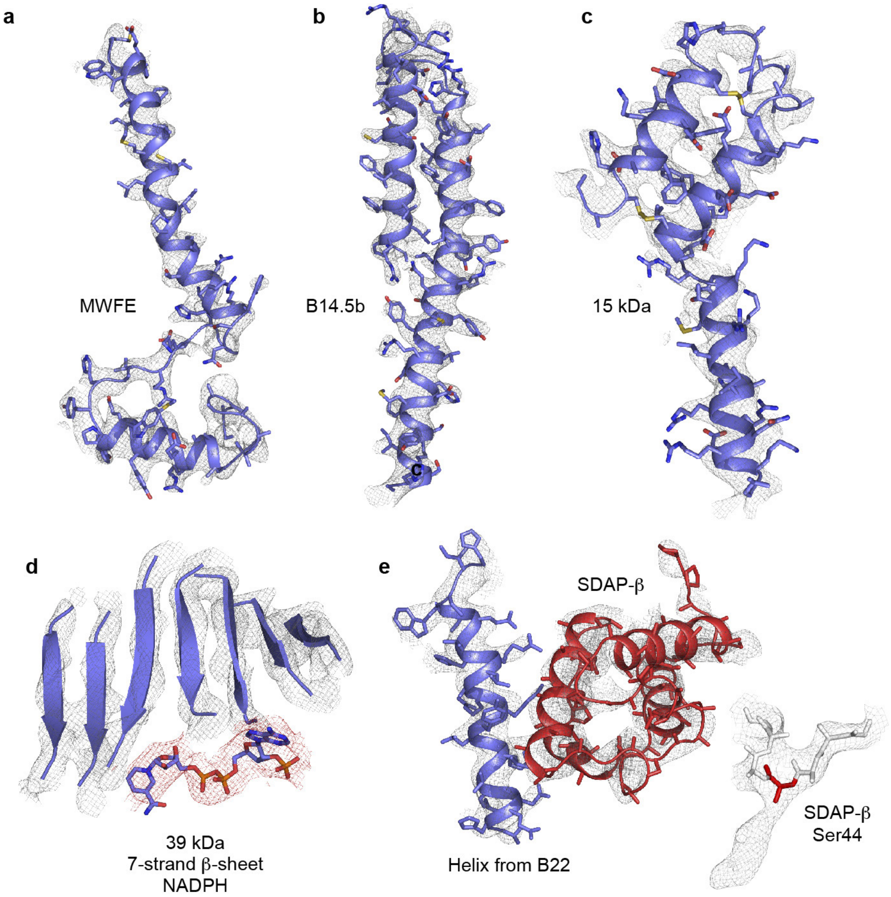
Extended Data Figure 1 | Resolution estimation and ResMap analysis of the density map for complex I before classification. a, The map, shown at two different contour levels, is coloured according to the local resolution, as determined by ResMap[35]. At the higher contour level (left), the majority of the protein is resolved to 3.9–4.7 Å; only the very peripheral regions (parts of the 51 kDa, 24 kDa and 75 kDa subunits in the matrix arm, and the distal end of the membrane arm) are at lower resolutions of 5–6 Å. At the lower contour level (right), the detergent/lipid belt in the 7–9 Å resolution range dominates the lower-resolution features. **b**, A slice through the map shows that large portions of the central, core regions are resolved to 4 Å or better. **c**, The FSC curve defines an estimated overall resolution of 4.16 Å at FSC = 0.143.



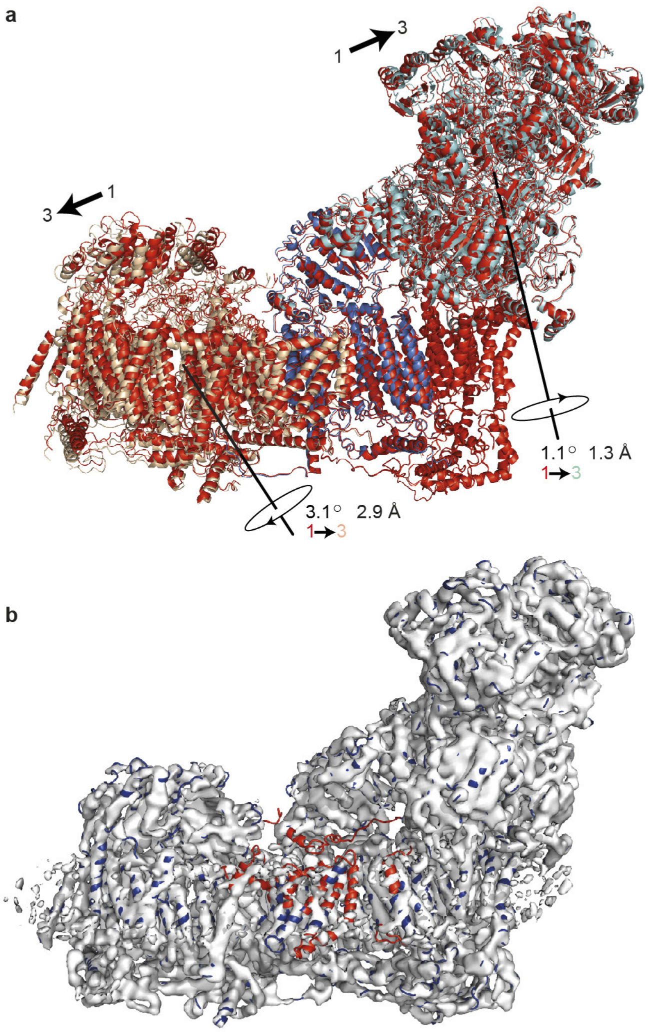
Extended Data Figure 2 | Resolution estimation and ResMap analysis of the classes of complex I. a–c, Data on classes 1, 2 and 3, respectively. Classes 1 and 2 display similar distributions in local resolution, with the majority of the protein in the range 4—5 Å. In class 3 the majority of protein displays a resolution of 4.5—5 Å. In all three cases the refined models agree very well with the maps as shown by comparison of the FSC curves (red) and the FSC curves from the half-maps (blue), and the similarity of the resolution values at FSC 0.143 and 0.5. The estimated resolutions, defined where the line at FSC = 0.143 crosses the blue curve, are 4.27 Å for class 1, 4.35 Å for class 2 and 5.6 Å for class 3. **d**, Cross-validation of the refinement parameters, confirming lack of over-fitting. For classes 1 and 2, one of the half maps was used for refinement then the FSC curves were calculated for each of the two half maps using the same model.



Extended Data Figure 3 | Example regions of the cryoEM density map for the core subunits, and the model fitted to the map. a, A selection of TMHs from the membrane domain: TMH3 from ND1, the distorted TMHs in ND6, ND4 and ND5, and a discontinuous TMH from ND2. The series of TMHs from left to right illustrates the decrease in resolution along the domain. **b**, The two [4FeS4] clusters in the 75 kDa subunit (density in red, at higher contour level) with the protein ligating one of them. **c**, The FMN cofactor in the 51 kDa subunit. **d**, The β-sheet in subunit PSST, showing clear separation between the strands. **e**, Two helices from the 49 kDa subunit.



Extended Data Figure 4 | Example regions of the cryoEM density map for the supernumerary subunits, and the model fitted to the map. **a**, Subunit MWFE, containing one TMH. **b**, Subunit B14.5, containing two TMHs. The N- and C-terminal loops are not shown. **c**, The 15 kDa subunit on the IMS face, containing a CHCH domain with two disulfide bonds. The N- and C-terminal loops are not shown. **d**, The seven-strand β-sheet in the 39 kDa subunit, showing the separation of the strands, and the bound nucleotide (red density) modelled as NADPH. **e**, Helix 1, one of the arginine-rich helices, in B22, and SDAP-β, on the matrix side of the tip of the membrane domain. Inset: the weak density attached to Ser44 in SDAP-β attributed to the attached pantetheine-4'-phosphate group (side chain of Ser44 not modelled).



Extended Data Figure 5 | Relationships between classes 1 and 3. a, The structures for classes 1 and 3 have been superimposed using ND1 and ND3. In class 3, relative to class 1, the hydrophilic and distal membrane domains are both rotated and shifted, but the change in the membrane domain dominates. Although the changes appear to make the angle of the L-shaped molecule increase they do not originate from a hinge-like motion at the interface of the hydrophilic and membrane domains. Class 1 is in red, class 3 is in red (ND1 domain), wheat, blue and cyan. Details of the composition and movement of the domains are given in Extended Data Table 4. **b**, The density for class 3 (white) is presented with the model for well-resolved regions of class 3 in blue (the model is enclosed in the density) and the model for poorly resolved regions in red (the model appears outside the density). The poorly resolved regions include the N terminus of the 49 kDa subunit and the transverse helix in ND5, as well as elements of ND4, ND6, B14.7, ESSS and B15.

Extended Data Table 1 | Summary of the models for the core subunits of *B. taurus* complex I

Subunit	Other names*	Chain	Total residues	Modelled residues	Assigned residues	Unknown residues	% residues modelled	% residues assigned	% with sidechains	% unknown residues
ND1 class 1	Nqo8 NuoH	H	318	3-200 218-315	3-200 218-315	-	93.1	93.1	89.0	0
ND1 class 2	Nqo8 NuoH	H	318	3-315	3-315	-	98.4	98.4	89.0	0
ND2	Nqo14 NuoN	N	347	2-345	2-345	-	99.1	99.1	87.6	0
ND3 class 1	Nqo7 NuoA	A	115	2-27 51-112	2-27 51-112	-	76.5	76.5	72.2	0
ND3 class 2	Nqo7 NuoA	A	115	2-112	2-112	-	96.5	96.5	72.2	0
ND4	Nqo13 NuoM	M	459	3-459	3-459	-	99.6	99.6	96.3	0
ND4L	Nqo11 NuoK	K	98	2-96	2-96	-	96.9	96.9	96.9	0
ND5	Nqo12 NuoL	L	606	2-605	2-605	-	99.7	99.7	88.1	0
ND6	Nqo10 NuoJ	J	175	2-172	2-172	-	97.7	97.7	79.4	0
75 kDa	NDUFS1 Nqo3 NuoG	G	704	8-692	8-209	210-692	97.3	28.7	8.9	68.6
51 kDa	NDUFV1 Nqo1 NuoF	F	444	14-438	14-438	-	95.7	95.7	18.2	0
49 kDa class 1	NDUFS2 Nqo4 NuoCD	D	430	5-50 61-430	5-50 61-430	-	96.7	96.7	87.0	0
49 kDa class 2	NDUFS2 Nqo4 NuoCD	D	430	5-430	5-430	-	99.1	99.1	89.3	0
30 kDa	NDUFS3 Nqo5 NuoCD	C	228	8-213	8-213	-	90.4	90.4	85.5	0
24 kDa	NDUFV2 Nqo2 NuoE	E	217	8-193	8-193	-	85.7	85.7	1.8	0
PSST	NDUFS7 Nqo6 NuoB	B	179	27-173	27-173	-	82.1	82.1	82.1	0
TYKY	NDUFS8 Nqo9 NuoI	I	176	1-176	1-176	-	100	100	92.0	0

* The names of the human, *T. thermophilus* and *E. coli* subunits (if different to the names in *B. taurus*).

Extended Data Table 2 | Summary of the models for the supernumerary subunits of *B. taurus* complex I

Subunit	Human name	Chain	Total residues	Modelled residues	Assigned residues	Unknown residues	% residues modelled	% residues assigned	% with sidechains	% unknown residues
42 kDa	NDUFA10	O	320	5-318	91-318	5-90	98.1	71.3	26.9	26.9
39 kDa class 1	NDUFA9	P	345	2-186 200-252 280-324	-	2-186 200-252 280-324	82.0	0	0	82.0
39 kDa class 2	NDUFA9	P	345	2-336	-	2-336	97.1	0	0	97.1
18 kDa	NDUFS4	Q	133	16-128	-	16-128	85.0	0	0	85.0
15 kDa	NDUFS5	e	105	6-94	6-94	-	84.8	84.8	41.9	0
13 kDa	NDUFS6	R	96	5-93	59-93	5-58	92.7	36.5	27.1	56.3
10 kDa	NDUFV3	s	75	1-35†	-	1-35†	46.7	0	0	46.7
AGGG	NDUFB2	j*	72	8-59	-	8-59	72.2	0	0	72.2
ASHI	NDUFB8	l	158	5-122	-	5-122	74.7	0	0	74.7
ESSS	NDUFB11	g	125	25-121	25-121	-	77.6	77.6	40.8	0
KFYI	NDUFC1	c	49	1-46	1-46	-	93.9	93.9	55.1	0
MNLL	NDUFB1	f	57	3-56	3-56	-	94.7	94.7	42.1	0
MWFE	NDUFA1	a	70	1-64	1-64	-	91.4	91.4	71.4	0
PDSW	NDUFB10	p	175	4-172	76-142	4-75 143-172	96.6	38.3	30.9	58.3
PGIV	NDUFA8	X	171	5-168	5-113	114-168	95.9	63.7	53.8	32.2
SDAP-α	NDUFAB1	T	88	8-82	8-82	-	85.2	85.2	0	0
SDAP-β	NDUFAB1	U	88	4-88	4-88	-	96.6	96.6	0	0
SGDH	NDUFB5	h	143	7-140	7-45	46-140	93.7	27.3	21.0	66.4
B22	NDUFB9	n	178	10-175	10-136	137-175	93.3	71.3	35.4	21.9
B18	NDUFB7	o	136	57-114	57-114	-	42.6	42.6	2.9	0
B17.2	NDUFA12	q	145	2-139	2-139	-	95.2	95.2	0	0
B17	NDUFB6	i	127	6-32 40-118	6-32	40-118	83.5	21.3	14.2	62.2
B16.6	NDUFA13	Z	143	5-142	31-99	5-30 100-142	96.5	48.3	48.3	48.3
B15	NDUFB4	m	128	11-128	32-128	11-31	92.2	75.8	66.4	16.4
B14.7	NDUFA11	Y	140	1-138	1-138	-	98.6	98.6	98.6	0
B14.5a	NDUFA7	r	112	4-70 100-119†	-	4-70 100-119†	77.7	0	0	77.7
B14.5b	NDUFC2	d	120	3-116	29-97	3-28 98-116	95.0	57.5	57.5	37.5
B14	NDUFA6	W	127	16-126	16-126	-	87.4	87.4	57.5	0
B13	NDUFA5	V	115	8-113	8-113	-	92.2	92.2	40.0	0
B12	NDUFB3	k*	97	1-74†	-	1-74†	76.3	0	0	76.3
B9	NDUFA3	b	83	1-80	1-45	46-80	96.4	54.2	54.2	42.2
B8	NDUFA2	S	98	16-95	16-95	-	81.6	81.6	0	0

*The assignments of B12 and AGGG may be reversed
†Arbitrary residue numbers

Extended Data Table 3 | Subunit–subunit interactions for the supernumerary subunits

	Primary core subunit	Subunit interactions	Notes
42 kDa	ND2	ND2, ND4, 49 kDa (N-terminus), ESSS, KFYI, B14.5b, B13	Nucleoside kinase fold; does not contact hydrophilic core subunits or B14+SDAP-α
39 kDa	PSST	ND1, ND3 (class 2 only), ND6, 75 kDa, 30 kDa, PSST, TYKY, 18 kDa, 13 kDa, B14	Short-chain dehydrogenase/reductase fold[20] with bound NAD(P)(H)[21]
18 kDa	75 kDa	75 kDa, 51 kDa, 49 kDa, 30 kDa, TYKY, 39 kDa, 10 kDa, B17.2, B14	Four-strand β-sheet with helix located between 75 and 30 kDa
15 kDa	ND2	ND2, ND3, ND4L, ND6, PGIV, SGDH, B16.6, B14.5b	CHCH domain; IMS[13]
13 kDa	TYKY	75 kDa, 49 kDa, TYKY, 39 kDa, B17.2	Zinc-binding domain[22] at the interface of TYKY, 75 and 49 kDa
10 kDa	24 kDa	75 kDa, 51 kDa, 24 kDa, 18 kDa	Present in the three-subunit flavoprotein subcomplex with 51 and 24 kDa
AGGG	ND5	ND5, B12	1 TMH; uncertain assignment of AGGG and B12 between chains j and k
ASHI	ND5	ND4, ND5, B22, B15	1 TMH; large globular domain on the matrix face, TMH crosses transverse helix
ESSS	ND4	ND4, ND5, 42 kDa, PDSW, SGDH, B22, B15, B14.5b	1 TMH; poorly resolved N-terminus on matrix face; possible disulphide 112 to PDSW 154.
KFYI	ND2	42 kDa, B14.5b	1 TMH; attached to complex by B14.5b
MNLL	ND4	ND4, PDSW, SGDH	1 TMH
MWFE	ND1	ND1, ND6, TYKY, PGIV, B17.2, B16.6, B14.5a	1 TMH; runs alongside ND1 TMH1 at proposed entrance to Q-binding site
PGIV	ND1	ND1, ND2, ND4, 15 kDa, MWFE, SGDH, B16.6, B14.5b, B9	Two CHCH domains; IMS[13]
PDSW	ND4	ND4, ND5, ESSS, MNLL, SGDH, B18, B17, B15, B14.5b	Extensive helix structure on IMS face; two likely internal disulphides (112-124, 76-83); possible disulphide 154 to ESSS-112
SDAP-α	30 kDa	B14	ACP on the hydrophilic domain[16]
SDAP-β	ND5	ND5, ASHI, B22, B17, B12	ACP on the membrane domain[16]
SGDH	ND4	ND2, ND4, ND5, ND6, 15 kDa, ESSS, MNLL, PDSW, PGIV, B22, B17, B16.6, B14.5b	1 TMH; long helix running along the IMS face of the membrane domain
B22	ND5	ND4, ND5, ASHI, SDAP-β, SGDH, B17, B15, B12	LYR protein that binds SDAP-β
B18	ND5	ND5, AGGG, PDSW, B17	CHCH domain; IMS[13]
B17.2	TYKY	ND1, 75 kDa, PSST, TYKY, 18 kDa, 13 kDa, MWFE, B14.5a	Three-strand β-sheet and long loop running across hydrophilic domain
B17	ND5	ND5, PDSW, SDAP-β, SGDH, B22	1 TMH; helix on the matrix face, β-strand in IMS augments β-hairpin between ND5 TMHs
B16.6	ND1	ND1, ND3, ND6, 49 kDa, TYKY, 15 kDa, MWFE, PGIV, SGDH, B14.5a, B9	1 TMH; 65-residue helix that crosses the membrane then turns along the IMS face
B15	ND4	ND4, ND5, ASHI, PDSW, B22, B14.7	1 TMH; long helix runs across the matrix face of the membrane domain
B14.7	ND2	ND2, ND4, ND5, SGDH, B15	4 TMHs; on the anchor of the transverse helix; likely disulphide 18-75
B14.5a	49 kDa	75 kDa, 51 kDa, 49 kDa, 30 kDa, TYKY, MWFE, B17.2, B16.6, B13	Long loop structure over hydrophilic domain
B14.5b	ND2	ND2, ND4, 42 kDa, 15 kDa, ESSS, KFYI, PDSW, PGIV, SGDH, B15	2 TMHs; anchors KFYI to complex
B14	30 kDa	75 kDa, 49 kDa, 30 kDa, PSST, 39 kDa, 18 kDa, SDAP-α,	LYR protein that binds SDAP-α[15]
B13	30 kDa	49 kDa, 30 kDa, 42 kDa, B14.5a	Three-helix bundle
B12	ND5	ND5, SDAP-β, AGGG, B22	1 TMH; uncertain assignment of AGGG and B12 between chains j and k
B9	ND1	ND1, ND3, ND6, TYKY, 15 kDa, PGIV, B16.6	1 TMH
B8	75 kDa	75 kDa	Thioredoxin-like fold[38]. Possible disulphide 23-57

Extended Data Table 4 | Allocation of subunits to domains, and the relative movement of domains between classes 1, 2 and 3

	Subunits	Transformation
Class 1 vs. class 2		
Heel domain	ND1, ND3 MWFE, B9, PGIV, B16.6	None (reference domain)
Membrane domain	ND2, ND4L, ND6, ND4, ND5, N-terminus of 49 kDa subunit* 42 kDa, 15 kDa, KFYI, B14.5b, B14.7, MNLL, AGGG, B12, B15, SGDH, B17, B18, ASHI, B22+SDAP-β, PDSW, ESSS	3.9° rotation 0.9 Å shift
Hydrophilic domain	75 kDa, 51 kDa, 24 kDa, 30 kDa, 49 kDa (except its N-terminus*), PSST, TYKY B8, B13, B14+SDAP-α, B14.5a, 39 kDa, B17.2, 18 kDa, 13 kDa, 10 kDa	3.4° rotation 2.3 Å shift
Class 1 vs. class 3		
Heel domain	ND1, ND3 MWFE, B9, PGIV, B16.6	None (reference domain)
Proximal membrane domain	ND2, ND4L, ND6 42 kDa, 15 kDa, KFYI, B14.5b	0.8° rotation 0.4 Å shift
Distal membrane domain	ND4, ND5 B14.7, MNLL, AGGG, B12, B15, SGDH, B17, B18, ASHI, B22+SDAP-β, PDSW, ESSS	3.1° rotation 2.9 Å shift
Hydrophilic domain	75 kDa, 51 kDa, 24 kDa, 30 kDa, 49 kDa (except its N-terminus), PSST, TYKY B8, B13, B14+SDAP-α, B14.5a, 39 kDa, B17.2, 18 kDa, 13 kDa, 10 kDa	1.1° rotation 1.3 Å shift

*The N terminus of the 49 kDa subunit (residues 5–39) is displaced in class 1 relative to class 2 when considered from the core fold of the subunit because it lies on the surface of the membrane domain and moves with ND2.

Extended Data Table 5 | Data collection, refinement and model statistics for classes 1 and 2

	Class 1	Class 2
Data collection		
Pixel size (Å)	1.33	1.33
Defocus range (µm)	1.8 - 5.5	1.8 – 5.5
Voltage (kV)	300	300
No. of particles	48,033	33,301
Orientation accuracy (º)	0.92	0.95
Model composition		
Non-hydrogen atoms	51,117	51,652
Protein residues	7,789	7,891
% of total	91.5	92.7
Core subunit residues	4,294	4,344
% of total	95.5	96.6
Supernumerary subunit residues	3,495	3,547
% of total	87.0	88.3
Refinement		
Resolution (Å)	4.27	4.35
Average B-factor (Å^2)	93.4	110.4
RMS deviations		
Bonds (Å)	0.008	0.008
Angles (°)	1.38	1.40
Validation		
Molprobity score	2.11	2.50
Clashscore, all atoms	2.91	3.35
Ramachandran plot		
Favoured (%)	86.66	86.92
Outliers (%)	3.57	3.72

CORRIGENDUM

doi:10.1038/nature17421

Corrigendum: CEACAM1 regulates TIM-3-mediated tolerance and exhaustion

Yu-Hwa Huang, Chen Zhu, Yasuyuki Kondo, Ana C. Anderson, Amit Gandhi, Andrew Russell, Stephanie K. Dougan, Britt-Sabina Petersen, Espen Melum, Thomas Pertel, Kiera L. Clayton, Monika Raab, Qiang Chen, Nicole Beauchemin, Paul J. Yazaki, Michal Pyzik, Mario A. Ostrowski, Jonathan N. Glickman, Christopher E. Rudd, Hidde L. Ploegh, Andre Franke, Gregory A. Petsko, Vijay K. Kuchroo & Richard S. Blumberg

Nature **517,** 386–390 (2015); doi:10.1038/nature13848

In this Letter, we published the crystal structure of a heterodimer of the human (h)CEACAM1 IgV domain and hTIM-3 IgV domain (Protein Data Bank (PDB) accession 4QYC). Since publication, E. Sundberg and S. Almo have questioned our model, and stated that they had obtained better results refining a hCEACAM1–hCEACAM1 homodimer model against our diffracted amplitudes. We confirm that a homodimer model indeed fits our crystallographic data better, as judged by most statistical measures (see Supplementary Table 1). We have therefore withdrawn the deposited heterodimer model (PDB code 4QYC) from the PDB, and replaced it with a more accurate homodimer model (PDB code 5DZL). We thank E. Sundberg and S. Almo for bringing this to our attention, and apologize for any confusion the original structure may have caused.

Our error was rooted in an assumption, which now seems to be invalid, that a crystal of a chimaeric two-domain protein should give X-ray diffraction data reflecting both domains. We sought evidence for the hTIM-3 domain in our data using multiple molecular replacement and crystallographic strategies, and also reprocessed the data using lower symmetry space groups in case the hTIM-3 signal was lost in the computational averaging. We now believe that the relatively low resolution of the dataset (3.4 Å) and the similarities between the folds of hCEACAM1 and hTIM-3 make it impossible to model the hTIM-3 IgV domain confidently using the data at hand. To understand how dimeric hCEACAM1 could predominate in a crystal built from a protein construct designed to ensure a 1:1 ratio of hCEACAM1 and hTIM-3, we performed western blot analyses on the materials used for crystallization, which showed a predominant species of ~26 kDa, consistent with intact chimaeric protein and minor additional lower molecular mass species, suggesting proteolysis as the reason for the absence of hTIM-3, consistent with the long time (months) required for crystal growth (Supplementary Fig. 1). Considering the strong tendency of CEACAM1 to crystallize as dimers, even a small amount of free homodimer could have preferentially crystallized.

In light of this, it is important to consider whether the incorrect structure calls into question any of the other results and conclusions of the paper. Presumably, the chimaeric protein used in our Letter was unaffected because the conditions did not favour proteolysis. Nevertheless, we felt obliged to extend our biophysical experiments to provide additional support for a direct interaction between hCEACAM1 and hTIM-3. Co-crystallization attempts with hCEACAM1 and hTIM-3 would be hindered by the tendency of CEACAM1 to homodimerize with a dissociation constant (K_d) of 450 nM (ref. 1). We instead pursued NMR and surface plasmon resonance (SPR) studies using purified tag-free IgV domains of hTIM-3 and hCEACAM1 (Supplementary Methods and Supplementary Fig. 2a–c). NMR ^{15}N-HSQC spectra of ^{15}N-labelled hCEACAM1 IgV domains showed spectral changes after incubation with unlabelled hTIM-3 IgV in the presence of 2 mM calcium (Supplementary Fig. 3a), which binds to the FG loop of hTIM-3 IgV (Supplementary Fig. 3b); calcium alone did not induce spectral changes in the ^{15}N-labelled hCEACAM1 ^{15}N-HSQC spectra (Supplementary Fig. 2d). So far, hCEACAM1 dimerization and oligomerization at experimental concentrations for NMR prevent us from mapping the spectral changes to the putative hTIM-3 binding site on the GFCC′ face of hCEACAM1. We also performed SPR experiments in which we immobilized hCEACAM1 IgV and flowed over hTIM-3 IgV plus calcium at varying concentrations. In the initial SPR experiments, 250 response units (RU) of hCEACAM1 IgV protein were directly immobilized via amine coupling and resulted in low levels of the hTIM-3 receptive surface ($B_{max} < 10\%$), probably due to misoriented and dimeric immobilized hCEACAM1. Global fit analysis of the hTIM-3 binding sensorgrams to a 1:1 Langmuir binding model (Supplementary Fig. 3c, top) yielded an underestimated association rate, an overestimated slow disassociation rate, and an overall K_d value of 2.2 μM (Supplementary Table 2). Further SPR studies used an oriented strategy in which C-terminal biotinylated hCEACAM1 IgV was immobilized at predominantly monomeric concentrations to a neutravidin-coupled biosensor chip, allowing CEACAM1 N-termini to be directed toward the solution. Single cycle kinetic studies were performed with serial hTIM-3 concentrations and both improved kinetic fit analysis and extrapolated steady-state analysis were performed using a 1:1 Langmuir binding model (Supplementary Fig. 3d and Supplementary Table 2), which yielded similar affinity constants (2–6 μM) to that obtained with the amine coupling experiments.

In conclusion, our new solution-based NMR and surface-based SPR studies provide further independent biophysical evidence to support a direct interaction between hCEACAM1 and hTIM-3 via their N-terminal IgV domains. After withdrawal of our crystallographic model, we cannot confidently state the stoichiometry, describe the molecular details, or differentiate between *cis*/*trans* modes of IgV domain interaction, as claimed in Letter. Future studies are needed to determine whether the interaction may be further facilitated by additional factors or by higher order oligomerization of hCEACAM1.

We also provide a corrected version of Fig. 2i (as Supplementary Fig. 3e to this Corrigendum), which during the review process inadvertently duplicated the left panel of the autoradiogram.

The following authors contributed to this Corrigendum: Andrew Russell, Zhen-Yu J. Sun, Walter Kim, Yasuyuki Kondo, Amit Gandhi, Yu-Hwa Huang, Daniel A. Bonsor, Sebastian Günther, Eric J. Sundberg, Vijay Kuchroo, Gerhard Wagner, Greg Petsko and Richard S. Blumberg.

Supplementary Information is available in the online version of the Corrigendum.

1. Bonsor, D., Gunther, S., Beadenkopf, R., Beckett, D. & Sundberg, E. Diverse oligomeric states of CEACAM IgV domains. *Proc. Natl Acad. Sci. USA* **112,** 13561–13566 (2015).

CORRIGENDUM

doi:10.1038/nature17985

Corrigendum: A receptor heteromer mediates the male perception of female attractants in plants

Tong Wang, Liang Liang, Yong Xue, Peng-Fei Jia, Wei Chen, Meng-Xia Zhang, Ying-Chun Wang, Hong-Ju Li & Wei-Cai Yang

Nature **531,** 241–244 (2016); doi:10.1038/nature16975

In Fig. 3f of this Letter the 'minus' symbol in the first column next to GST–MK1^{KD} should be a 'plus'. In addition, the labels 'At*DIS1*' in Fig. 4d should read 'At*MDIS1*'. These errors have been corrected online.

CORRIGENDUM

doi:10.1038/nature18279

Corrigendum: Mitochondrial ROS regulate thermogenic energy expenditure and sulfenylation of UCP1

Edward T. Chouchani, Lawrence Kazak, Mark P. Jedrychowski, Gina Z. Lu, Brian K. Erickson, John Szpyt, Kerry A. Pierce, Dina Laznik-Bogoslavski, Ramalingam Vetrivelan, Clary B. Clish, Alan J. Robinson, Steve P. Gygi & Bruce M. Spiegelman

Nature **532,** 112–116 (2016); doi:10.1038/nature17399

In this Letter, owing to a typographical error, Fig. 4h was erroneously referred to as Fig. 4j in the text and in the Fig. 4 legend. This error has been corrected online.

CORRIGENDUM

doi:10.1038/nature18281

Corrigendum: Mycocerosic acid synthase exemplifies the architecture of reducing polyketide synthases

Dominik A. Herbst, Roman P. Jakob, Franziska Zähringer & Timm Maier

Nature **531,** 533–537 (2016); doi:10.1038/nature16993

In this Letter, we studied the three-dimensional structure of a protein from *Mycobacterium smegmatis* assigned as mycocerosic acid synthase (MAS) in sequence databases as A0R1E8 in Uniprot (http://www.uniprot.org/uniprot/A0R1E8) and YP_888986.1 in NCBI (https://www.ncbi.nlm.nih.gov/protein/118473069). In conclusion, we provided a template structure of MAS-like polyketide synthases (PKSs) and a first example of reducing PKS architecture. However, we now note that Etienne *et al.*[1] provided a biochemical characterization of a deletion strain of the corresponding gene MSMEG_4727 (https://www.ncbi.nlm.nih.gov/gene/4534621), which indicated a physiological role of the protein in the production of 2,4-dimethyl-2-eicosenoic acid, a lipid component of lipooligosaccharides, rather than mycocerosic acids, via a reaction closely related to those of MAS. Until comprehensive characterization at the protein level is available, the protein we studied should therefore be referred to as a 'mycocerosic-acid synthase like-PKS' or 'MAS-like PKS'; the database records will be updated accordingly. We thank the authors of ref. 1 for drawing our attention to this publication. The main scientific conclusions of our manuscript remain unchanged.

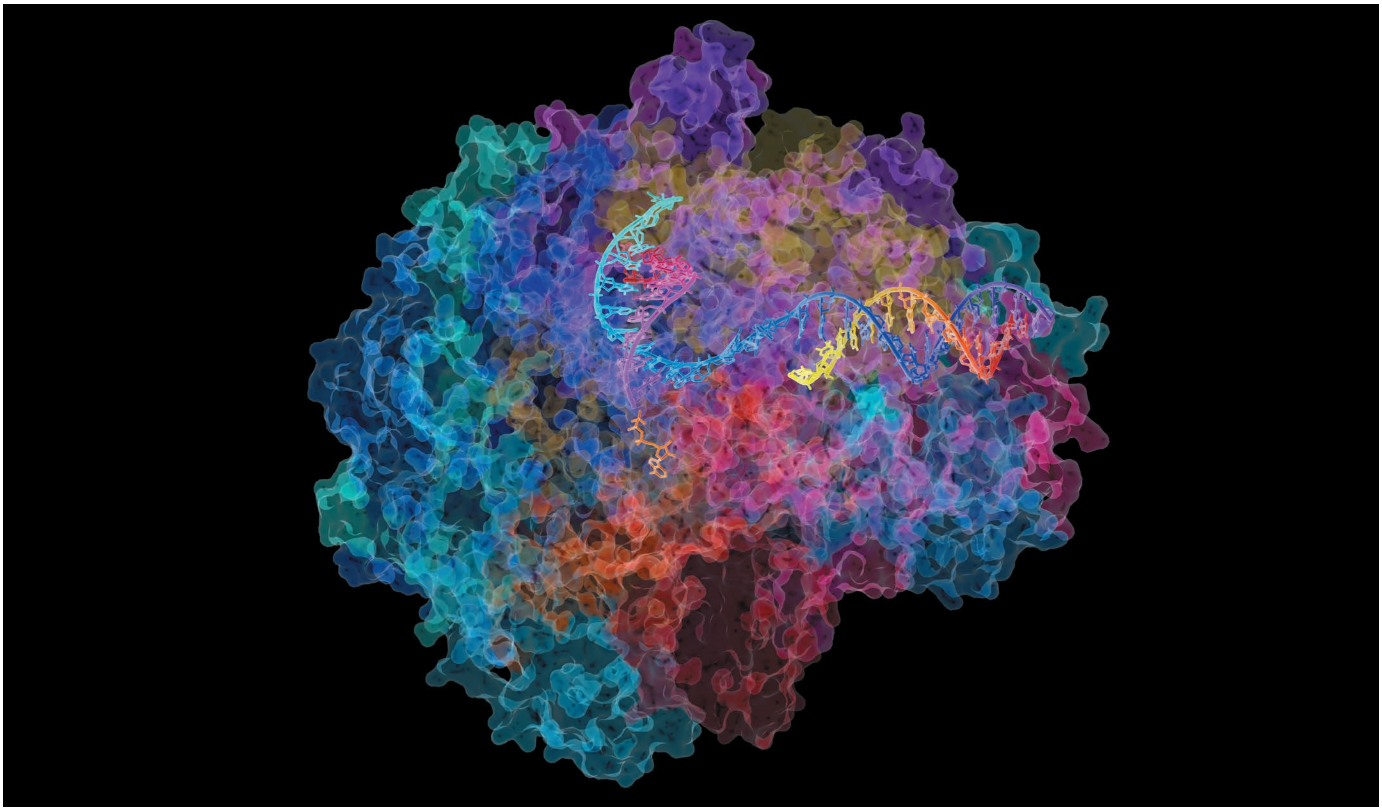
1. Etienne, G. *et al.* Identification of the polyketide synthase involved in the biosynthesis of the surface-exposed lipooligosaccharides in mycobacteria. *J. Bacteriol.* **191,** 2613–2621 (2009).

TECHNOLOGY FEATURE

LET THE STRUCTURAL SYMPHONY BEGIN

Structural biologists are at last living the dream of visualizing macromolecules to uncover their function. But it means integrating different technologies, and that's no easy feat.

DAVID BUSHNELL, KEN WESTOVER & ROGER KORNBERG/STANFORD UNIV.



The details of the enzyme RNA polymerase have been intriguing structural biologists for some time.

BY STEPHEN ORNES

Like other structural biologists, Eva Nogales works in extraordinary times. The University of California, Berkeley, faculty member now has the tools to tackle important questions about cells' molecular machinery that would have been impossible to answer just a few years ago.

A recent project with Berkeley colleague Jennifer Doudna, the molecular biologist who co-pioneered the CRISPR–Cas9 gene-editing method, is a case in point. Both were intensely interested in the R-loop, a structure made of nucleic acids that forms in cells in many situations, but also just before DNA is snipped by CRISPR–Cas9. Nogales and her team revealed an R-loop in *Streptococcus pyogenes* bacteria, and from the near-atomic-resolution images, deduced how the Cas9 enzyme opens up the DNA conformation at specific sites and makes them accessible to CRISPR's molecular scissors[1].

The work is remarkable for the speed with which the scientists assigned a function to the structure, but also because they arrived at the solution by combining imaging methods — an increasingly popular approach in structural biology. For more than a century, the field's premiere method has been X-ray crystallography. But some biomolecules are simply too big or small to crystallize, and the technique doesn't work on others. And some biomolecules change shape or orientation as they work, which isn't captured by static crystallization.

Now, scientists have a dazzling suite of different imaging techniques with which to build on crystallographic findings. Some of the approaches, such as cryogenic electron microscopy (cryo-EM) or chemists' stalwart nuclear magnetic resonance (NMR) imaging, reveal molecular shapes, size and orientation at near-atom-level resolution without the need to make crystals. But not every method works for every protein, nucleic acid or other biomolecule inside a living cell.

Growing wisdom in the field suggests that no single method is likely to be sufficient to probe the dynamic behaviour or intricate interactions taking place in a cell. The most powerful insights will come from hybrid ▶

methodologies that integrate the images from several different tools.

The approach is rapidly gaining followers. “Each [method] brings something important to the table, and the combination is very much larger than the sum of the parts,” structural biologist Roger Kornberg of Stanford University in California. Kornberg won the 2006 Nobel Prize in Chemistry for his work detailing the machinery of gene transcription. For that ground-breaking research, he generated crystallographic pictures. Now, like other crystallographers, he has moved on to hybrid methodologies.

Kornberg continues to analyse RNA polymerase II, but now he combines crystallography with cryo-EM, in which an electron beam probes the structure of biomolecules. Cryo-EM can be used on molecules that don’t crystallize easily and can reveal larger structures than can X-ray crystallography, but — for the moment at least — it lacks crystallography’s high resolution. Kornberg’s lab also uses chemical crosslinking and mass spectroscopy to reveal relationships between nearby proteins, and homology modelling to construct representations using information from known proteins[2].

Nogales and Doudna’s team also took the hybrid route to study R-loops. “The full R-loop could not be seen by the high-resolution X-ray crystallographic structure,” says Nogales. So they also used cryo-EM to reveal the full R-loop structure at lower resolution. Only by combining the two methods could the researchers work out how R-loops fit into the larger CRISPR–Cas9 picture[1].

Such hybrid, or integrative, approaches help researchers to probe deep basic-science questions, but also reveal details that are useful to drug developers. Large proteins found in cell membranes are often targets for therapeutic drugs, and high-resolution hybrid methods have the potential to show in atomic detail how a drug interacts with a receptor. Similarly, hybrid methods might be able to aid vaccine development by showing how proteins on the viral envelope of HIV, Ebola and other pathogens interact with immune cells to induce protective responses. “These structures are super-important to understand how our immune system works,” says structural biologist Jens Meiler at Vanderbilt University in Nashville, Tennessee.

Nogales sums it up: “This is a golden time to do hybrid methodologies.”

LIVING THE DREAM

The current era in structural biology promises to fulfil the “dream of many life scientists”, says Jan Ellenberg, head of the Cell Biology and Biophysics Unit at the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. That dream is to seamlessly scale up from what scientists see at the atomic level to the cellular level. Such deep understanding of the cell’s macromolecules naturally leads to answers to the overarching question in structural biology — how is a molecule’s structure connected to its function?

Each technique in a structural biologist’s toolbox offers a different perspective. Models that use hybrid methods can boost biologists’ confidence that a model accurately reflects how the molecule or ensemble acts in the cell. “You need all of them in combination to really get a full understanding of your biological question of interest,” says Meiler.

“One of the really exciting things about cryo-EM is that you can start a biochemical process and freeze those samples at multiple states.”

X-ray crystallography has long reigned as the standard way to determine the atomic structure of proteins. Of the 120,000 or so models in the Protein Data Bank (PDB), established in 1971, about 90% were derived from crystallographic studies.

But structural biologists’ workhorse, even with its high resolution, has limitations. Crystallography requires highly purified samples that produce a well-ordered crystal. Scientists fire X-rays at a crystal to determine its structure by analysing how the atoms scatter light. The technique needs a specimen with enough atoms to produce a measurable diffraction pattern, and every crystal must be static. As a result, the method can’t reveal how a molecule moves or functions in a cell, or its connections to other systems.

A protein “is not just a single static structure”, says Gunnar Schröder, who leads the computational structural biology group at the Institute of Complex Systems in Jülich, Germany. “Oftentimes, what you want is to see how the whole protein works.” Schröder uses hybrid methods to understand the movements and connections of proteins. Crystallography provides a snapshot of a protein in one configuration, removed from its normal environment. He says that structural biologists need other methods to boost the structural information from crystallography and improve their understanding of the form and function of proteins.

Many proteins, such as drug targets on a cell membrane, are flexible and often unstable. To get these proteins to form crystals, researchers often have to change them in some way. Meiler says the altered specimen may not accurately reflect the native state of the molecule or how it is arranged in the cell. He mixes experimental and computational approaches to better understand molecular structure. “It takes time for people to understand that for many biological systems, the model from crystallography is a good starting point,” he says, but it may not be suitable for providing information about function.

Biologists are now leveraging a range of tools to build richer, more accurate models of biological structures. Hybrid approaches have the power to do more than a single technique ever could. One particularly useful partnership joins cryo-EM and X-ray crystrallography. This microscopy method has been around since the 1980s, but in recent years it has achieved a resolution of 2.2 ångströms, edging close to the 2 Å average resolution of X-ray crystallography[3]. It can produce models in two or three dimensions of proteins and other macromolecules that have stubbornly resisted other approaches.

“One of the really exciting things about cryo-EM is that you can start a biochemical process and freeze those samples at multiple states,” says Jeffrey Lengyel, principal scientist for life sciences at FEI, a company in Hillsboro, Oregon, that designs and manufactures cryo-electron microscopes. “You can determine the structure of multiple conformations.”

Researchers can also combine cryo-EM images to see molecules in motion. John Rubinstein of the Hospital for Sick Children in Toronto, Canada, led work published in May 2015 that used image analysis to combine 100,000 cryo-EM images into a film, showing changes in the structure of eukarotic V-ATPase, an enzyme that pumps proteins across membranes and, changes over time[4]. In papers published earlier this year, Nogales and her collaborators used cryo-EM with homology models to describe the structure of TFIID, a large, horseshoe-shaped protein complex that is required to initiate gene transcription[5].

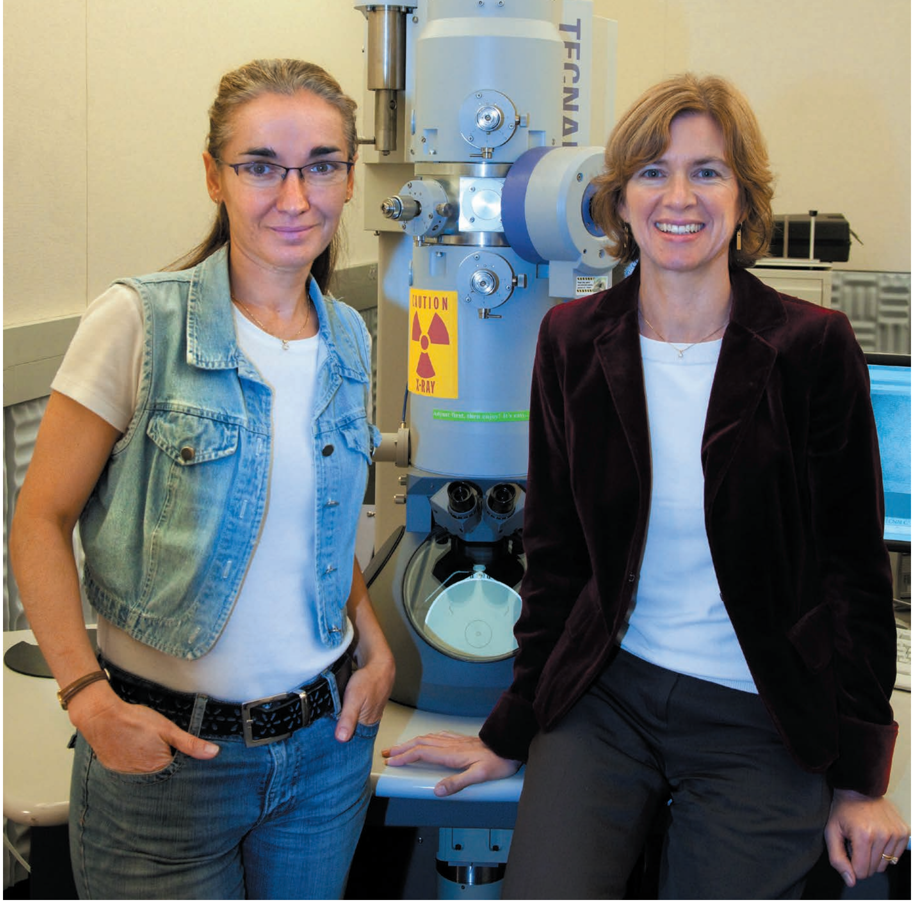
WITH A LITTLE HELP

The hybrid strategy of Nogales and her team led to an overall resolution of better than 10 Å — a significant improvement over their previous analysis of the same protein at 30 Å. That resolution has led to new insights: “We can see what amino acids are interacting with DNA,” Nogales says.

But cryo-EM requires specimens to be snap-frozen. That’s not ideal for biological samples, as the conditions are far removed from a macromolecule’s dynamic, natural state. NMR spectroscopy can help on that front. “NMR has a big advantage in that you can look at proteins at room temperature, and get information on dynamics,” says Schröder, whose lab builds experimental models that combine NMR data with those from cryo-EM and crystallography.

First used experimentally in the 1940s, NMR reveals macromolecular structures by exciting atoms in an external magnetic field. When the atoms relax, the changes in their internal magnetic fields can be mapped to each atom. However, NMR spectroscopy works only on relatively small macromolecules or ensembles.

Structural biologists are also using hybrid



LAWRENCE BERKELEY NATL LAB.

Eva Nogales (left) and Jennifer Doudna worked together to reveal how the Cas9 enzyme uncoils DNA in preparation for gene editing.

methods to tackle supersized ensembles, a task that would have been impossible in the past. Kornberg's latest research, which has not yet been published, extends his ongoing RNA polymerase II studies and uses hybrid methodologies to describe a giant assembly made of more than 50 proteins and transcription factors. "The entire assembly could now be visualized for the first time through the combination, and I would say equal contribution, of all the methods," he says.

Another supersized target is the nuclear pore complex. This collection of membrane proteins acts as a gatekeeper for information and molecules passing in and out of the nucleus. In 2015, Ellenberg and his colleagues used a hybrid approach to study the structure of this protein behemoth[6]. In the past, researchers had probed the complex with crystallography and electron microscopy, but they weren't able to image the entire thing at molecular resolution, and its overall structure largely remained a mystery.

Ellenberg's team first imaged the nuclear-pore complex using super-resolution fluorescence microscopy, which he says can identify features measuring less than 30 nanometres. To improve the resolution, they combined it with an image-processing technique called single-particle averaging that uses information from thousands of pores, bringing down the resolution to about 10 Å. Comparisons with cryo-EM maps of the same complex validated their work. The result is a zoomed-in view of a supersized protein complex. The EMBL team "generated models of the nuclear pore that were unthinkable in the past", Nogales says.

Similarly, Rubinstein and Lewis Kay at the University of Toronto used hybrid methods to push the boundaries of what was deemed possible. By combining cryo-EM with NMR spectroscopy, they mapped previously unidentified conformational changes of an enzyme called VAT, which has an important role in breaking down proteins in a cell. Cryo-EM revealed the structure, and used together with NMR, they were able to show how the enzyme changes shape, painting an elegant portrait of a protein at work[7].

HYBRID DRAWBACKS

Although biologists are gaining clarity from merging different tools, each technique also contributes its own error rates. Mixing them, therefore, presents a potential problem because it multiplies the sources of error. "How can I combine these different ways of analysing error into one holistic approach that gives me a measure of confidence, accuracy and precision in model?" asks Meiler.

Yet another hurdle is melding different data sets to make them accessible and useful to other researchers. The rich level of information from any one technique makes this a formidable challenge. "You can literally generate terabytes of data per day," says Lengyel. He hopes the structural-biology community might benefit from the approaches in astronomy and high-throughput genetics to grapple with data overload. Although software exists that can neatly combine high-resolution crystallographic data into cryo-EM maps, other hybrid methodologies aren't as straightforward to merge. Electron paramagnetic resonance spectroscopy, for example, measures distances and orientation in a macromolecule, whereas cryo-EM produces a density map. Although those two measurements would be useful together, they don't speak the same language. "How do I combine these very different metrics? How do I share these data?" asks Meiler.

To discuss the best ways to organize, share and use data from hybrid approaches, dozens of structural biologists gathered in October 2014 at the European Bioinformatics Institute in Hinxton, UK[8]. The meeting was the first of its kind, organized by a task force set up by the Worldwide Protein Data Bank. At present, the PDB stores data from individual protein structures, says Schröder. "We should get to the point where we have all the information that we know about this protein — all the different conformations it can take," he says. Such rich data, he says, will help to reveal the bigger picture of proteins and other big molecules.

There have been steps in that direction: archives exist for electron-microscopy models in two dimensions (the Electron Microscopy Pilot Image Archive) and three dimensions (EMDataBank). Established with funding from the EMBL and other sources, these archives contain data that can be shared, archived and distributed.

Yet another challenge threatens to forestall progress in the field: human expertise. "Investments are needed in technology, but it's equally important to invest in educating scientists," says Meiler. He recommends that students learn the limitations and challenges of each method — and become an expert in at least one. "We need to train a new generation of scientists who are capable of understanding how to integrate these different technologies," he says.

Finally, structural biologists have to learn to ask new, complicated biological questions that may seem impossible. Thanks to hybrid methods, says Ellenberg, "things have come within reach that even five years ago I wouldn't have been able to dream of doing until retirement". ■

Stephen Ornes *is a science writer in Nashville, Tennessee.*

1. Jiang, F. *et al. Science* **351,** 867–871 (2016).
2. Murakami, K. *et al. Proc. Natl Acad. Sci. USA* **112,** 13543–13548 (2015).
3. Bartesaghi, A. *et al. Science* **348,** 1147–1151 (2015).
4. Zhao, J., Benlekbir, S. & Rubinstein, J. L. *Nature* **521,** 241–245 (2015).
5. Louder, R. K. *et al. Nature* **531,** 604–609 (2016).
6. Szymborska, A. *et al. Science* **341,** 655–658 (2013).
7. Huang, R. *et al. Proc. Natl Acad. Sci. USA* **113,** E4190–E4199 (2016).
8. Sali, A. *et al. Structure* **23,** 1156–1167 (2015).

CHOOSING DIRECTION From grad student to blog poster **go.nature.com/2alz8oq**

COMMUNICATION Sharing science with the public **go.nature.com/2ba0ks3**

NATUREJOBS For the latest career listings and advice **www.naturejobs.com**

IMAGENAVI/GETTY

Researchers should learn the local custom for introducing themselves to colleagues.

WORK ABROAD

Visa to visit

Researchers working outside their home country should be careful to brush up on local customs.

BY BARBRA RODRIGUEZ

Travel abroad as a graduate student, postdoc or visiting researcher can be daunting. You may be blindsided by cultural values and behavioural protocols that are vastly different from those of your home country, and snags may arise over tacit conventions on what to say and how to act, both in and out of the workplace.

Preparation is key if you want to avoid gaffes and missteps. So before you leave, consult websites and books such as the *Culture Shock!* series (published by Marshall Cavendish). That way, you'll be able to familiarize yourself with the mores of your host nation and learn, for example, how people greet one another (see 'Proper introductions'), manage conflict and share lab resources.

Once abroad, chat often with colleagues and others, both in the lab and outside, and socialize whenever possible. This will help your colleagues warm to you. It can also help you to distinguish individual personality traits from broad cultural tendencies. Whatever your approach, making an effort to respect your host country's culture builds goodwill. "They don't mind if you make mistakes," says Adrian Moore, a British team leader at the RIKEN Brain Science Institute (BSI) in Saitama, Japan. "There's one rule for foreigners and another for the Japanese."

Many nations, particularly those in northern Europe, use direct communication styles, which can feel jarring to people from other cultures. Amanda Henry, a physical anthropologist from the United States, was taken aback by frank comments from German colleagues at the Max Planck Institute for Evolutionary Anthropology in Leipzig. Technicians in her lab told her candidly about the 30 or more attempts they had already made to identify an experimental contaminant. "It was, 'We tried this and it didn't work. We tried this and it didn't work,' over and over, without trying to sugarcoat it or focus on fixing it, as Americans would do," Henry says. Her words of encouragement were met with stiffened bodies and a reply that they weren't upset. "I had totally read an emotional response into something that was entirely rote for them," she says. "Now I try to hear what is there, rather than interpret — and I usually ask for feedback."

Seeking the opinions of others becomes especially important in cultures where people tend to communicate indirectly in the lab. Despina Goniotaki, a neuroscience PhD student from Greece, learnt that she had to deliberately reach out to Swiss colleagues at the University Hospital Zurich for advice. They were hesitant to give potentially negative yet possibly crucial feedback. "I presented my first cell-surface biotinylation study during an official lab meeting, but I had to arrange a meeting afterwards with a postdoc about what else to try."

EMOTIVE DIVIDE

Direct communication is one thing; expression of emotion is another. Lisandra Zepeda, a bioinformatics graduate student at the University of Copenhagen who is from Mexico, recalls her surprise when a Dane who had lost essential computer data merely frowned and said: "OK, I'll think about what to do next." "Somebody in Mexico would scream 'Ah!', but she just kept quiet," Zepeda says.

Reserved cultures can seem unfriendly to researchers from countries where people wear their feelings on their sleeves. Goniotaki, who is finishing her fifth year in Switzerland, says the lack of reaction when she grumbled or talked excitedly about experiments during breaks has meant that she speaks up less nowadays. "I would get indirect comments about how sometimes I would overreact about things that lab mates considered not so important," ▶

she says. She has developed friendships with Swiss people, but prefers to use exercise or playing the piano when she feels the need to vent frustrations.

Finding the right outlet to meet personal and professional needs also helps when dealing with challenges that relate to hierarchy. Developmental neuroscientist Douglas Campbell at the RIKEN BSI recalls that a graduate student's practice talk for a thesis defence livened up after the principal investigator left. Until that point, his colleagues had felt inhibited because of Japan's emphasis on seniority. "The lab members then shared their opinions for several hours," says Campbell, who is now a visiting scientist at the Technical University of Munich.

Other potential stumbling blocks include expectations about boundaries — both literal and figurative — between personal and professional life. Kelsey Glennon, a US population geneticist at the University of the Witwatersrand in Johannesburg, South Africa, found that she had to ask students from indigenous tribes not to stand so close to her. "They're good at accommodating for my American sense of personal space and laugh about this," she says. Shira Raveh-Rubin, an Israeli postdoc at the Swiss Federal Institute of Technology Zurich, was taken aback when she mentioned her children to lab mates and an awkward silence ensued. She eventually found some Swiss colleagues who were more receptive to family-oriented conversations, and who shared their own stories when she mentioned family matters. "I found that if I don't open up, then I become depressed," she says. "You have to respect others, but still really be yourself."

Cultural differences may be much less obvious, and awkward, in a large international lab: expectations for newcomers may be looser, and colleagues more accommodating, than is often the case in a lab where most people belong to the same culture. Dulce Vargas Landín, a Mexico City native and PhD student at the University of Western Australia in Perth, is now on her third overseas venture; her current lab mates hail from Croatia, India, Vietnam, Italy and Brazil, among other places. She says that their diverse approaches balance out individual differences, such as those relating to the expression of strong feelings. "It's a good result when the lab is heterogeneous," she says. "Not too much drama and you get work done."

START EARLY

Junior scientists can evaluate potential labs through study-abroad programmes such as EuroScholars, or through summer programmes or internships at institutions such as the RIKEN BSI and the Technion Israel Institute of Technology in Haifa. Graduate students, postdocs and visiting researchers can also find out beforehand whether their prospective principal investigator has worked abroad — a plus. "[Those who have] learn how to deal with different personalities and cultural backgrounds," Vargas Landín says. For instance, during a lab presentation, her Australian leader redirected questions from a German colleague that Vargas Landín suspects were becoming too narrowly focused.

> *"You break the ice by going out drinking and to dinner. That's where they test you to find out who you really are."*

Once you're in your new country, it can be useful to have a cultural adviser, such as a language coach, who can help you to understand what sorts of issues could affect your relationships, and what people are likely to expect of you. Your institution may set you up with such an adviser in your host nation. You should also aim to befriend a local colleague who can fill that role. But if you find yourself clashing with someone, it's best to speak first with the person directly concerned instead of turning to other people. That's the advice of Anne Copeland, a clinical psychologist in Boston, Massachusetts, who helps those who live and work in unfamiliar cultures. "We jump to blaming a person's character," she says. "That's too bad, because often it's a learned cultural difference."

Socializing with colleagues outside the lab can often provide insight into how people interact and help you to understand how, and when, to recalibrate your behaviour. For example, you might join in at the tea breaks that serve as social glue at many research institutions in the United Kingdom, Australia and South Africa. Jonah Choiniere, a palaeontologist from the United States who is a faculty member at the University of the Witwatersrand, initially struggled with the idea of stopping work for morning teas. But he has come to appreciate them in the past four years. "It is an effective way to disseminate information and remain on good footing with your colleagues," he says.

Shared lunch hours and breaks are commonplace in France, Australia and elsewhere, and after-hours gatherings occur regularly in Asian and other nations that favour formal work relationships. "You break the ice by going out drinking and to dinner," says Moore, in Japan. "That's where they test you to find out who you really are." He says that colleagues and senior researchers are more likely to tell you in these settings if there is something they are unhappy with you about — but that none of this should be discussed back at the lab.

The reward for putting up with cultural friction away from home can be a more accepting approach to others when you return. These experiences can strengthen interpersonal and leadership skills. "Before I started travelling, I would have just gotten mad whenever there was a lab conflict," Vargas Landín says. "Now, if someone is from a different culture and thinks there's a better way to do an experiment, I can give it a try. And maybe we'll discover a new way to do things in the end." ■

Barbra Rodriguez *is a freelance writer in Austin, Texas.*

RELATIONSHIP BUILDING

Proper introductions

Greetings can vary greatly between cultures, says Terri Morrison, lead author of the book *Kiss, Bow or Shake Hands* (Adams Media, 2006), which provides business travellers with cultural overviews and relationship tips for more than 60 nations. Here are a few pointers to starting relationships well.

- **Formality.** In Germany, China, Japan and elsewhere, you should address a supervisor or colleague by surname only for some time. A good approach is to wait for an invitation to use a first name.
- **Personal information.** As with surnames, use comments from colleagues to guide how much to share about your private life. Natives of some host nations may be looking for shared interests. If so, you might mention possibilities such as parenthood, sports or a local delicacy.
- **Handshakes.** Those in Asian and other nations such as South Africa may prefer gentle handshakes. In Japan, it is helpful to stand far enough away from the person you're greeting, especially in formal settings, to allow room for a bow (with eyes cast downward). A careful exchange of business cards may follow. Researchers from the United Kingdom, northern Europe or the United States, where a firm handshake and steady eye gaze suggest reliability, should remember that these behaviours can signal aggressiveness in Asian and other nations. In South Africa, a more extended type of handshake is common among some cultures, such as the Sutu and Zulu tribes. Tribal members may change hand positioning, and continue to hold hands after the handshake .
- **Interaction focus.** The United States and nations in northern Europe value workplace accomplishments and productivity, so researchers from those regions tend to use succinct greetings. Conversely, those from Latin America, Greece and the Middle East, among other regions, often view greetings as a step towards building a relationship that helps to establish your credibility. They prefer a conversation to a quick handshake, so making time for that is important. B.R.

LEGACY ADMISSIONS

A degree of uncertainty.

BY S. R. ALGERNON

ILLUSTRATION BY JACEY

Mr Lindstrom shook my hand. It was a firm six-fingered grip. I could feel the bony nub where the seventh finger had been. Lindstrom looked around the empty banqueting hall and signed his name at the registration table by the door. The banner over his head read ALUMNI NETWORKING DINNER.

"Let me guess," I said. "Class of sixty-seven?"

"Sixty-three, actually."

His parents must have been early adopters. Maybe that explained the rebellious amputation. Fourteen-fingered kids had been all the rage back in the Sixties, in homage to the keyboardist for The Rolling Beagles, a centenary tribute band to the British Invasion. That genetic tweak summed up — for me, at least — how our parents had got it all wrong. It wasn't their fault, though. What else could they have done, really?

"I'm glad you could make it," I said. "There are complimentary drinks at the bar, and the food will be ready soon. On your registration card, you listed 'Biometric consultant' as your occupation. That sounds exciting."

"Yeah," he said. He picked his name tag out from all the others and fidgeted with it. "I work for a bank. Transactions are automated now, so after a while the biometric scanner starts to forget what a real human looks like in the flesh. I help it to recalibrate its sensors. To tell you the truth, I think it's just lonely."

"Oh," I said. "If you aren't happy with your current job, we can help you find another. The university prides itself on a 100% employment rate for its graduates. That's a lifetime guarantee."

"I don't know yet. I think we need to get away from here and find someplace by the ocean, someplace with … I don't know … trees and everything, like it used to be."

He withdrew from our conversation as a memory flashed in his eyes with stark clarity.

"Eidetic?" I asked. He nodded. Photographic memories were all the rage, too, along with various learning and attention enhancements. The university recruited only the best applicants, with impeccable test scores and clear academic potential.

"Sorry. Is there anything I can do?"

"I don't think so," he said, with a hint of shame breaking through the nostalgia.

"Do you have any children?"

"Janice and Todd," he said, but without the pride that people used to have when they talked about such things.

"Don't worry," I said. "We have a legacy programme, you know. We'll have no problem getting them financial assistance, if they're keeping their test scores up."

"Not much point in that anymore."

"Well, is there any way I can help? It is my job after all."

"I don't know really," he said, glancing back at the doorway. "I guess I was just curious. I wanted to see if the place was still here, you know. I thought there might be people I could talk to."

"It's still early yet." I looked around at the circular banquet tables, each with six flawless place settings. The automated caterer rolled between them, with food trays in tow. "At least you could stay for the buffet."

"Sorry," he said. "I should be going. Still, it was nice meeting you."

"It was good of you to come. Please do remember our legacy programmes. The university is here for you."

"Thank you," he said, "I'll keep that in mind."

If I have to, was the unspoken addendum, *if the ocean and the trees no longer sustain us.*

Mr Lindstrom left. I was alone again, except for the machines that gathered up the food for reclamation promptly at nine.

At least someone showed up, I thought. *Better than last year.*

➲ NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

An empty self-driving shuttle van waited for me in the parking lot, the same one that had deposited me three hours earlier. It said nothing, and I was in no mood to break the silence.

I cut through the history department building on the way to the administration building. The cleaning bot whirred cheerfully as I walked past rows of offices. Each door was diligently locked, but if the university saw fit to let me inside, I was sure that I would find the offices in pristine condition, waiting for the next faculty hire.

"Good evening," I said to Professor Emeritus Franklin, through the open doorway to her office. She looked up from a paperback book and smiled at me as I passed.

Back at the administration building, on the first floor, my office occupied the end of the hall. The sign on the door read OFFICE OF RECRUITMENT AND RETENTION. Its magnetic lock opened as I approached.

Once inside, I passed the empty cubicles and took a seat in the corner office. My three doctorates hung on the wall, along with the handful of master's degrees.

"University," I said, "could you add Todd and Janice Lindstrom to our list of potential recruits for …?"

"2097 and 2099?"

"Yes, that's right," I sighed. "Is there anything else on my calendar for the week?"

"The next alumni event is in three months, twelve days. In the meantime, there are 2,793 online courses that you are qualified for, and as always you are entitled to a staff tuition waiver. Shall I enrol you in a course with one of our virtual professors?"

"Not tonight."

I cracked all twelve of my knuckles and swivelled my chair for a view of the vestigial skyline.

Were there people behind those windows, I wondered, or did the buildings leave the lights on to remind themselves how things used to be? I took some comfort in knowing that we humans weren't the only ones still caught up in the past. ■

S. R. Algernon *studied fiction writing and biology, among other things, at the University of North Carolina at Chapel Hill. He currently lives in Singapore.*